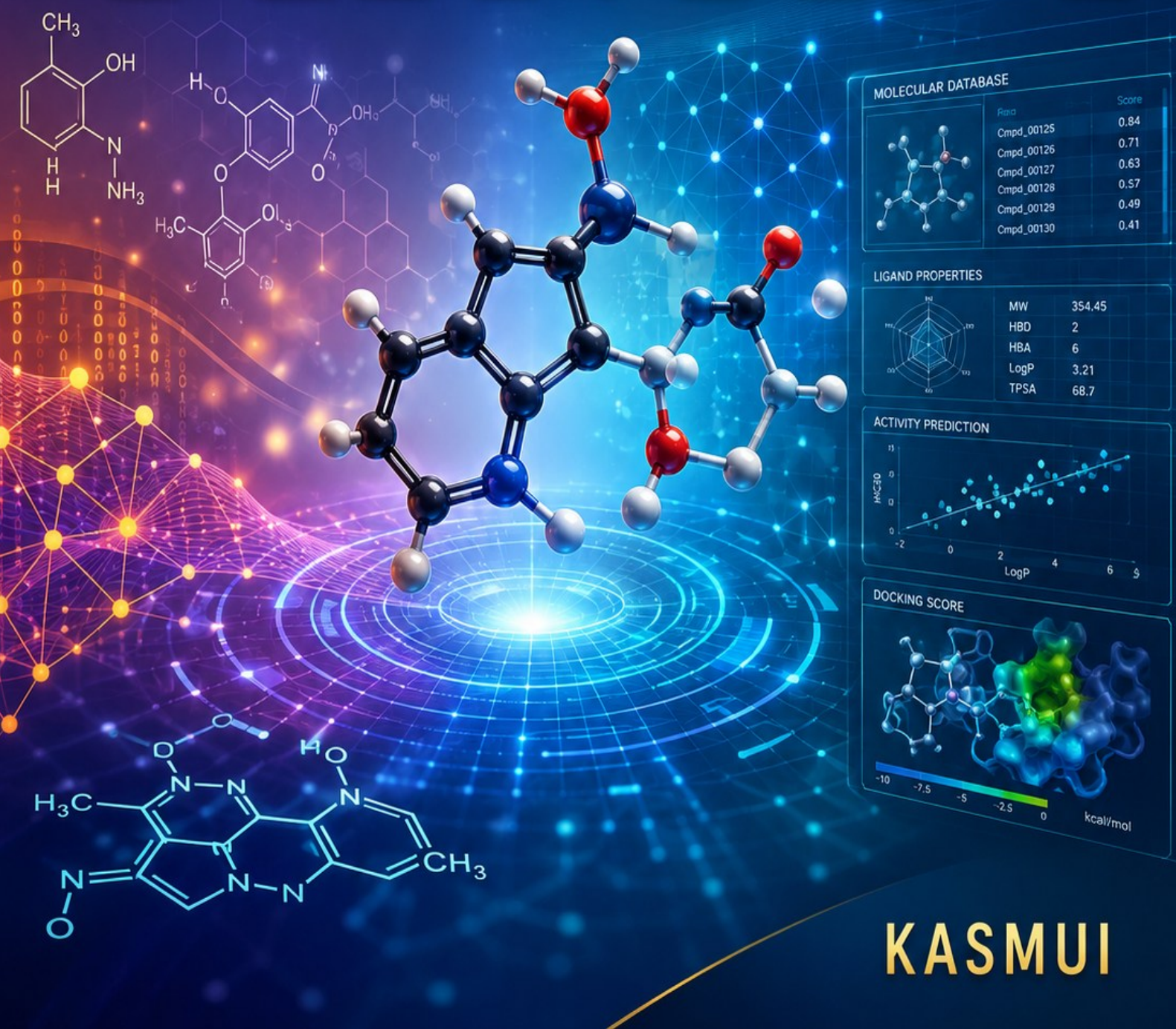
INFORMATIKA KIMIA



KASMUI


INFORMATIKA KIMIA
Representasi Molekul, Data Kimia, Pemodelan,
Pembelajaran Mesin, dan Penemuan Berbasis Komputer
KASMUI
Buku Ajar untuk Mahasiswa Kimia S1

INFORMATIKA KIMIA

Representasi Molekul, Data Kimia, Pemodelan, Pembelajaran Mesin, dan Penemuan Berbasis Komputer

Buku Ajar untuk Mahasiswa Program Sarjana (S1) Kimia

Disusun oleh
KASMUI

Semarang, 2026

CATATAN PENERBITAN DAN PENGGUNAAN

Naskah ini merupakan buku ajar orisinal berbahasa Indonesia yang disusun sebagai sintesis konseptual dari dua sumber utama: An Introduction to Chemoinformatics karya Andrew R. Leach dan Valerie J. Gillet, serta Chemoinformatics: Theory, Practice, and Products karya Barry A. Bunin, Jürgen Bajorath, Brian Siesel, dan Guillermo Morales. Struktur, contoh, penjelasan, ilustrasi, latihan, dan kode dalam buku ini disusun ulang untuk konteks pembelajaran mahasiswa kimia S1; buku ini bukan terjemahan atau reproduksi kedua sumber tersebut.

Nama perangkat lunak, basis data, standar, dan produk merupakan hak pemilik masing-masing. Pembaca wajib memeriksa lisensi, versi, dokumentasi, dan ketentuan penggunaan terkini sebelum menggunakannya dalam riset atau layanan. Semua prediksi komputasi harus diperlakukan sebagai dukungan keputusan yang memerlukan validasi kimia dan eksperimen.

Ilustrasi dalam buku ini dibuat secara orisinal untuk naskah ini. Contoh kode bersifat edukatif dan perlu diuji kembali pada lingkungan pengguna. Dataset yang digunakan dalam tugas mahasiswa harus memiliki izin penggunaan yang jelas dan tidak memuat data pribadi atau rahasia tanpa persetujuan yang sah.

KATA PENGANTAR

Perkembangan kimia modern ditandai oleh pertumbuhan data yang sangat cepat. Struktur molekul, hasil spektroskopi, reaksi, bioaktivitas, sifat fisikokimia, dokumen paten, dan catatan laboratorium kini tersedia dalam bentuk digital dengan skala yang tidak dapat lagi dikelola hanya melalui pembacaan manual. Informatika kimia hadir sebagai disiplin yang menghubungkan pengetahuan kimia dengan representasi komputer, basis data, statistika, algoritme, dan kecerdasan buatan.

Buku ini disusun untuk mahasiswa kimia S1 agar mampu memahami bukan hanya cara menggunakan perangkat, tetapi juga alasan ilmiah di balik setiap representasi dan metode. Mahasiswa diajak melihat molekul sebagai objek kimia sekaligus objek informasi. Setiap keputusan komputasi—mulai dari penulisan SMILES, pemilihan deskriptor, penentuan koefisien kemiripan, pembagian data, sampai validasi model—dapat mengubah interpretasi kimia. Oleh sebab itu, buku ini menekankan kurasi, provenance, validasi, domain aplikasi, ketidakpastian, dan reprodusibilitas.

Materi disusun dalam sepuluh bab. Dua buku rujukan klasik menjadi tulang punggung konseptual: satu memberikan uraian sistematis tentang representasi 2D dan 3D, deskriptor, QSAR, kemiripan, keragaman, HTS, virtual screening, dan desain pustaka; sumber lain memperluas perspektif teori, praktik, produk, basis data, dan penerapan dalam penemuan senyawa. Buku ini mengintegrasikan keduanya serta menambahkan perkembangan mutakhir seperti prinsip FAIR, API, RDKit dan Python, graph neural network, model generatif, informatika reaksi, open science, dan paket riset reprodusibel.

Setiap bab memuat capaian pembelajaran, konsep inti, ilustrasi, persamaan, studi kasus, aktivitas, praktikum, latihan, dan refleksi. Lampiran menyediakan panduan proyek penulisan buku mini, tugas kelompok maksimal tiga mahasiswa, rubrik penilaian, format laporan, checklist kualitas data, glosarium A–Z, serta daftar pustaka pelengkap. Struktur tersebut diharapkan mendukung pembelajaran berbasis proyek dan mendorong mahasiswa menghasilkan portofolio ilmiah yang dapat diperiksa.

Ucapan terima kasih disampaikan kepada para penulis sumber utama, komunitas perangkat lunak terbuka, pengelola basis data kimia, serta para peneliti yang membangun fondasi informatika kimia. Kritik dan koreksi ilmiah terhadap naskah ini sangat diharapkan karena bidang informatika kimia berkembang cepat. Semoga buku ini membantu mahasiswa menghubungkan ketelitian kimia, penalaran matematis, dan praktik komputasi yang bertanggung jawab.

Semarang, Juli 2026 — Kasmui

CARA MENGGUNAKAN BUKU

Buku ini dapat digunakan sebagai bahan kuliah dua semester, buku referensi mandiri, atau dasar pembelajaran berbasis proyek. Untuk satu semester, dosen dapat memilih Bab 1–7 sebagai fondasi dan menggunakan Bab 8–10 sebagai topik aplikasi. Untuk dua semester, seluruh bab dapat disertai praktikum dan proyek buku mini.

Komponen	Fungsi	Cara menggunakan
Capaian pembelajaran	Menjelaskan kompetensi yang dituju	Gunakan sebagai kontrak belajar dan dasar asesmen.
Istilah kunci	Membentuk kosakata disiplin	Buat kartu istilah dan hubungan antarkonsep.
Ilustrasi dan tabel	Meringkas relasi data dan metode	Jelaskan kembali tanpa membaca caption.
Praktikum	Mengubah teori menjadi prosedur	Kerjakan berkelompok maksimal tiga mahasiswa.
Latihan	Menguji pemahaman dan evaluasi	Jawab dengan argumen, data, atau kode.
Proyek buku mini	Mengintegrasikan bab menjadi produk akademik	Gunakan rubrik pada lampiran.

CAPAIAN PEMBELAJARAN BUKU

- Menjelaskan ruang lingkup, sejarah, etika, dan siklus hidup data informatika kimia.
- Merepresentasikan dan memanipulasi struktur molekul 2D/3D, reaksi, serta metadata.
- Merancang basis data, melakukan kurasi, dan menggunakan identifier serta API secara bertanggung jawab.
- Menghitung dan mengevaluasi deskriptor, fingerprint, kemiripan, keragaman, dan ruang kimia.
- Membangun serta memvalidasi model QSAR/QSPR dan pembelajaran mesin tanpa data leakage.
- Menyusun alur penapisan virtual, analisis HTS, docking, dan prediksi ADMET yang dapat diaudit.
- Mendesain pustaka kombinatorial, menganalisis reaksi, serta memahami model generatif dan retrosintesis.
- Membuat pipeline Python/RDKit yang terdokumentasi, diuji, dan dapat direproduksi.
- Mengomunikasikan hasil melalui laporan, gambar, tabel, model card, data card, dan buku mini.

DAFTAR SINGKATAN

Singkatan	Kepanjangan
ADMET	Absorption, Distribution, Metabolism, Excretion, and Toxicity
API	Application Programming Interface
ECFP	Extended-Connectivity Fingerprint
ELN	Electronic Laboratory Notebook
FAIR	Findable, Accessible, Interoperable, Reusable
GNN	Graph Neural Network
HTS	High-Throughput Screening
InChI	International Chemical Identifier
LIMS	Laboratory Information Management System
MCS	Maximum Common Substructure
PCA	Principal Component Analysis
QSAR	Quantitative Structure–Activity Relationship
QSPR	Quantitative Structure–Property Relationship
RDF	Resource Description Framework
RMSD	Root-Mean-Square Deviation
SAR	Structure–Activity Relationship
SDF	Structure Data File
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SVM	Support Vector Machine
TPSA	Topological Polar Surface Area
UMAP	Uniform Manifold Approximation and Projection
VS	Virtual Screening

DAFTAR ISI

CATATAN PENERBITAN DAN PENGGUNAAN 4
KATA PENGANTAR 5
CARA MENGGUNAKAN BUKU 6
CAPAIAN PEMBELAJARAN BUKU 7
DAFTAR SINGKATAN 8
DAFTAR ISI 9
DAFTAR GAMBAR 29
BAB 1. LANDASAN INFORMATIKA KIMIA DAN EKOSISTEM DATA 30
Capaian Pembelajaran Bab 30
Istilah Kunci 30
1.1 Definisi, Ruang Lingkup, dan Identitas Disiplin 31
Pendalaman I — Objek, Makna, dan Tingkat Representasi 33
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 34
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 36
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan37
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 39
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 40
1.2 Sejarah Perkembangan dan Akar Keilmuan 41
Pendalaman I — Objek, Makna, dan Tingkat Representasi 43
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 45
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 46
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan48
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 49
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 51
1.3 Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data 52
Pendalaman I — Objek, Makna, dan Tingkat Representasi 54
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 56
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 57
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan59
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 60
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 62
1.4 Spektrum Data Kimia dan Siklus Hidup Data 63
Pendalaman I — Objek, Makna, dan Tingkat Representasi 65
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 67
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 68
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan70
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 71
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 73

1.5 Prinsip FAIR, Provenance, dan Interoperabilitas 74
Pendalaman I — Objek, Makna, dan Tingkat Representasi 76
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja .. 78
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 79
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan81
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 82
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 84
1.6 Etika, Lisensi, Privasi, dan Integritas Ilmiah 85
Pendalaman I — Objek, Makna, dan Tingkat Representasi 87
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja .. 88
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 90
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan91
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 93
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 94
1.7 Kompetensi Inti Mahasiswa Kimia S1 95
Pendalaman I — Objek, Makna, dan Tingkat Representasi 98
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja .. 99
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 101
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan102
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 104
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 105
1.8 Kerangka Pemecahan Masalah Informatika Kimia 106
Pendalaman I — Objek, Makna, dan Tingkat Representasi 108
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 110
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 111
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan113
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 114
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 116
Praktikum 1 — Audit Siklus Hidup Data Kimia 117
Tujuan 117
Bahan dan Perangkat 117
Langkah Kerja 117
Luaran Wajib 118
Pertanyaan Analisis 118
Ringkasan Bab 118
Latihan dan Tugas Individu 118
Refleksi Bab 120
BAB 2. REPRESENTASI DAN MANIPULASI STRUKTUR MOLEKUL 2D 121
Capaian Pembelajaran Bab 121
Istilah Kunci 121

2.1 Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi 122
Pendalaman I — Objek, Makna, dan Tingkat Representasi 124
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 125
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 127
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 128
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 130
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 131
2.2 Tabel Konektivitas dan Format Berkas 133
Pendalaman I — Objek, Makna, dan Tingkat Representasi 134
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 136
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 137
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 139
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 140
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 142
2.3 Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey 143
Pendalaman I — Objek, Makna, dan Tingkat Representasi 145
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 147
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 148
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 150
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 151
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 153
2.4 Aromatisitas, Valensi, Muatan, Tautomer, dan Protonasi 154
Pendalaman I — Objek, Makna, dan Tingkat Representasi 156
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 158
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 159
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 161
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 162
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 163
2.5 Stereokimia dan Isomerisme dalam Representasi Digital 165
Pendalaman I — Objek, Makna, dan Tingkat Representasi 167
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 169
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 170
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 172
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 173
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 174
2.6 Isomorfisme Graf dan Pencarian Substruktur 176
Pendalaman I — Objek, Makna, dan Tingkat Representasi 178
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 179

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 181
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan182
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 184
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 185
2.7 Basis Data Reaksi dan Representasi Transformasi 186
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 188
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 190
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 191
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan193
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 194
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 196
2.8 Paten, Struktur Markush, dan Tantangan Ambiguitas 197
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 199
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 200
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 202
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan203
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 205
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 206
Persamaan dan Metrik Penting.....207
Praktikum 2 — SMILES, SMARTS, dan Pencarian Substruktur.....208
Tujuan..... 208
Bahan dan Perangkat 208
Langkah Kerja 208
Luaran Wajib 208
Pertanyaan Analisis 209
Ringkasan Bab.....209
Latihan dan Tugas Individu.....209
Refleksi Bab 210
BAB 3. STRUKTUR 3D, KONFORMASI, FARMAKOFOR, DAN VISUALISASI 211
Capaian Pembelajaran Bab.....211
Istilah Kunci 211
3.1 Koordinat 3D dan Sumber Struktur Eksperimental 212
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 214
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 216
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 217
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan219
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 220
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 221
3.2 Geometri Molekul, Koordinat Internal, dan Orientasi 223

Pendalaman I — Objek, Makna, dan Tingkat Representasi 225
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 226
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 228
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan229
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 231
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 232
3.3 Energi Potensial dan Ruang Konformasi 233
Pendalaman I — Objek, Makna, dan Tingkat Representasi 236
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 237
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 239
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan240
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 242
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 243
3.4 Algoritme Pencarian Konformer 244
Pendalaman I — Objek, Makna, dan Tingkat Representasi 247
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 248
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 250
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan251
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 253
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 254
3.5 Penyelarasan, RMSD, dan Kemiripan Bentuk 255
Pendalaman I — Objek, Makna, dan Tingkat Representasi 258
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 259
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 261
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan262
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 264
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 265
3.6 Farmakofor dan Kueri 3D 266
Pendalaman I — Objek, Makna, dan Tingkat Representasi 268
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 270
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 271
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan273
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 274
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 275
3.7 Basis Data 3D Fleksibel dan Pencarian 277
Pendalaman I — Objek, Makna, dan Tingkat Representasi 279
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 280

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 282
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan284
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 285
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 286
3.8 Visualisasi Molekuler sebagai Alat Analisis.....288
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 290
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 291
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 293
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan294
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 296
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 297
Persamaan dan Metrik Penting.....298
Praktikum 3 — Generasi dan Analisis Konformer 298
Tujuan.....298
Bahan dan Perangkat 299
Langkah Kerja 299
Luaran Wajib 299
Pertanyaan Analisis 299
Ringkasan Bab.....300
Latihan dan Tugas Individu.....300
Refleksi Bab 301
BAB 4. BASIS DATA KIMIA, PENCARIAN, KURASI, DAN INTEROPERABILITAS 302
Capaian Pembelajaran Bab.....302
Istilah Kunci 302
4.1 Model Data Relasional dan Normalisasi 303
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 305
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 307
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 308
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan310
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 311
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 312
4.2 Basis Data Graf, Dokumen, dan Kolom.....314
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 316
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 317
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 319
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan320
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 322
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 323
4.3 Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri 324

Pendalaman I — Objek, Makna, dan Tingkat Representasi 326
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 328
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 329
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan331
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 332
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 334
4.4 Standardisasi, Deduplikasi, dan Rekonsiliasi Identitas 335
Pendalaman I — Objek, Makna, dan Tingkat Representasi 337
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 339
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 340
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan342
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 343
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 345
4.5 Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian 346
Pendalaman I — Objek, Makna, dan Tingkat Representasi 348
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 350
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 351
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan353
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 354
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 356
4.6 Ontologi, Taksonomi, dan Data Tertaut 357
Pendalaman I — Objek, Makna, dan Tingkat Representasi 359
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 360
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 362
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan363
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 365
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 366
4.7 Sumber Data Publik dan Komersial 367
Pendalaman I — Objek, Makna, dan Tingkat Representasi 370
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 371
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 373
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan374
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 376
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 377
4.8 API, Layanan Web, dan Otomasi Akuisisi Data 378
Pendalaman I — Objek, Makna, dan Tingkat Representasi 380
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 382

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 383
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan385
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 386
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 387
Praktikum 4 — Basis Data SQLite untuk Senyawa dan Assay 389
Tujuan..... 389
Bahan dan Perangkat 389
Langkah Kerja 389
Luaran Wajib 389
Pertanyaan Analisis 390
Ringkasan Bab.....390
Latihan dan Tugas Individu.....390
Refleksi Bab 391
BAB 5. DESKRIPTOR MOLEKULER, FINGERPRINT, DAN RUANG KIMIA 392
Capaian Pembelajaran Bab.....392
Istilah Kunci 392
5.1 Konsep Deskriptor dan Dimensi Representasi 393
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 395
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 397
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 398
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan400
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 401
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 403
5.2 Deskriptor Konstitusional dan Fisikokimia.....404
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 406
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 407
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 409
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan410
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 412
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 413
5.3 Indeks Topologi dan Deskriptor Graf 414
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 417
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 418
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 420
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan421
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 422
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 424
5.4 Fingerprint Struktural dan Circular Fingerprint 425
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 427

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 429
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 430
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan432
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 433
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 435
5.5 Deskriptor 3D, Medan, dan Permukaan 436
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 438
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 440
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 441
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan443
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 444
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 445
5.6 Praproses Data: Scaling, Transformasi, dan Missing Value 447
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 449
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 450
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 452
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan453
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 455
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 456
5.7 Korelasi, Seleksi Fitur, dan Reduksi Dimensi.....457
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 460
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 461
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 463
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan464
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 466
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 467
5.8 Visualisasi Ruang Kimia dan Batas Interpretasi 468
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 470
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 472
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 473
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan475
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 476
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 477
Persamaan dan Metrik Penting.....479
Praktikum 5 — Deskriptor, PCA, dan Peta Ruang Kimia 479
Tujuan..... 479
Bahan dan Perangkat 479
Langkah Kerja 479

Luaran Wajib 480
Pertanyaan Analisis 480
Ringkasan Bab 480
Latihan dan Tugas Individu 480
Refleksi Bab 481
BAB 6. KEMIRIPAN, KEANEKARAGAMAN, KLASTERING, DAN SELEKSI SENYAWA 483
Capaian Pembelajaran Bab 483
Istilah Kunci 483
6.1 Prinsip Sifat Serupa dan Activity Cliff 484
Pendalaman I — Objek, Makna, dan Tingkat Representasi 486
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 488
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 489
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 491
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 492
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 494
6.2 Koefisien Kemiripan untuk Fingerprint 495
Pendalaman I — Objek, Makna, dan Tingkat Representasi 497
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 498
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 500
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 501
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 503
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 504
6.3 Jarak untuk Deskriptor Kontinu 505
Pendalaman I — Objek, Makna, dan Tingkat Representasi 508
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 509
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 511
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 512
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 513
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 515
6.4 Kemiripan Substruktur, MCS, dan Reduced Graph 516
Pendalaman I — Objek, Makna, dan Tingkat Representasi 519
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 520
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 522
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan 523
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 524
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 526
6.5 Klastering Hierarkis dan Nonhierarkis 527
Pendalaman I — Objek, Makna, dan Tingkat Representasi 529
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 531

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 532
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan534
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 535
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 537
6.6 Pemilihan Set Beragam 538
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 540
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 541
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 543
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan544
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 546
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 547
6.7 Desain Set Latih, Validasi, dan Uji 548
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 551
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 552
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 554
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan555
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 557
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 558
6.8 Keanekaragaman Multi-Kriteria dan Keputusan Praktis.....559
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 561
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 563
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 564
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan566
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 567
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 568
Persamaan dan Metrik Penting.....570
Praktikum 6 — Kemiripan, Klastering, dan Seleksi Beragam.....570
Tujuan..... 570
Bahan dan Perangkat 570
Langkah Kerja 570
Luaran Wajib 571
Pertanyaan Analisis 571
Ringkasan Bab.....571
Latihan dan Tugas Individu.....572
Refleksi Bab 573
BAB 7. QSAR, QSPR, PEMBELAJARAN MESIN, DAN VALIDASI MODEL 574
Capaian Pembelajaran Bab.....574
Istilah Kunci 574
7.1 Perumusan Endpoint dan Desain Studi Prediktif 575

Pendalaman I — Objek, Makna, dan Tingkat Representasi 577
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 579
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 580
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan582
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 583
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 584
7.2 Regresi Linear, MLR, dan Dasar QSAR Klasik 586
Pendalaman I — Objek, Makna, dan Tingkat Representasi 588
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 589
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 591
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan592
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 593
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 595
7.3 PCR, PLS, Regularisasi, dan Seleksi Variabel 596
Pendalaman I — Objek, Makna, dan Tingkat Representasi 598
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 600
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 601
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan603
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 604
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 606
7.4 Klasifikasi: Logistik, Pohon, Random Forest, dan Boosting 607
Pendalaman I — Objek, Makna, dan Tingkat Representasi 609
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 611
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 612
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan614
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 615
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 617
7.5 SVM, Kernel, dan Metode Berbasis Tetangga 618
Pendalaman I — Objek, Makna, dan Tingkat Representasi 620
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 622
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 623
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan625
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 626
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 627
7.6 Jaringan Saraf, GNN, dan Representasi Terpelajar 629
Pendalaman I — Objek, Makna, dan Tingkat Representasi 631
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 632

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 634
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan635
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 637
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 638
7.7 Validasi Internal, Eksternal, dan Nested Cross-Validation.....639
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 641
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 643
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 644
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan646
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 647
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 649
7.8 Metrik Regresi, Klasifikasi, Ranking, dan Kalibrasi.....650
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 652
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 653
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 655
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan656
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 658
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 659
7.9 Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution ...661
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 663
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 664
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 666
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan667
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 669
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 670
7.10 Interpretabilitas, Kausalitas, dan Pelaporan Model.....672
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 673
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 675
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 677
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan678
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 680
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 681
Persamaan dan Metrik Penting.....682
Praktikum 7 — Model QSAR yang Divalidasi.....682
Tujuan..... 683
Bahan dan Perangkat 683
Langkah Kerja 683
Luaran Wajib 683

Pertanyaan Analisis 684
Ringkasan Bab..... 684
Latihan dan Tugas Individu..... 684
Refleksi Bab 685
BAB 8. HIGH-THROUGHPUT SCREENING, PENAPISAN VIRTUAL, DOCKING, DAN ADMET 686
Capaian Pembelajaran Bab..... 686
Istilah Kunci 686
8.1 Data High-Throughput Screening dan Kontrol Kualitas..... 687
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 689
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 691
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 692
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan694
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 695
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 696
8.2 Visualisasi dan Penambangan Hit 698
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 699
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 701
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 702
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan704
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 705
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 707
8.3 Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan..... 708
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 710
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 712
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 713
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan715
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 716
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 718
8.4 Ligand-Based Virtual Screening 719
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 721
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 723
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 724
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan726
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 727
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 729
8.5 Structure-Based Virtual Screening dan Preparasi Sistem 730
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 732
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 734

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 735
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan737
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 738
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 740
8.6 Algoritme Docking dan Sampling Pose 741
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 743
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 744
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 746
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan747
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 749
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 750
8.7 Fungsi Skor, Consensus Scoring, dan Rescoring 751
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 754
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 755
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 757
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan758
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 760
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 761
8.8 Prediksi ADMET dan Toksisitas.....762
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 764
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 766
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 767
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan769
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 770
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 771
8.9 Validasi Prospektif dan Integrasi Eksperimen 773
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 775
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 777
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 778
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan780
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 781
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 782
Persamaan dan Metrik Penting.....784
Praktikum 8 — Pipeline Penapisan Virtual Sederhana 784
Tujuan..... 784
Bahan dan Perangkat 784
Langkah Kerja 784
Luaran Wajib 785

Pertanyaan Analisis 785
Ringkasan Bab..... 785
Latihan dan Tugas Individu..... 785
Refleksi Bab 787
BAB 9. KIMIA KOMBINATORIAL, DESAIN PUSTAKA, INFORMATIKA REAKSI, DAN OPTIMASI..... 788
Capaian Pembelajaran Bab..... 788
Istilah Kunci 788
9.1 Prinsip Kimia Kombinatorial dan Enumerasi 789
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 791
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 793
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 794
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan796
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 797
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 798
9.2 Pustaka Beragam dan Pustaka Terfokus 800
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 802
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 803
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 805
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan806
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 808
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 809
9.3 Seleksi Monomer dan Seleksi Berbasis Produk..... 810
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 812
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 814
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 815
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan817
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 818
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 820
9.4 Desain Multiobjektif dan Front Pareto..... 821
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 823
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 825
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 827
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan828
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 830
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 831
9.5 Informatika Reaksi dan Prediksi Hasil..... 832
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 834
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 836

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 837
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan839
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 840
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 842
9.6 Retrosintesis Berbantuan Komputer 843
Pendalaman I — Objek, Makna, dan Tingkat Representasi 845
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 846
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 848
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan849
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 851
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 852
9.7 Model Generatif Molekul dan Kendala Kimia 854
Pendalaman I — Objek, Makna, dan Tingkat Representasi 856
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 857
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 859
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan860
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 862
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 863
9.8 Otomasi Laboratorium, ELN, LIMS, dan Design-Make-Test-Analyze864
Pendalaman I — Objek, Makna, dan Tingkat Representasi 866
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 868
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 869
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan871
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 872
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 874
Persamaan dan Metrik Penting 875
Praktikum 9 — Enumerasi Pustaka dan Optimasi Pareto 875
Tujuan 875
Bahan dan Perangkat 875
Langkah Kerja 876
Luaran Wajib 876
Pertanyaan Analisis 876
Ringkasan Bab 877
Latihan dan Tugas Individu 877
Refleksi Bab 878
BAB 10. PERANGKAT LUNAK, PEMROGRAMAN, REPRODUSIBILITAS, DAN ARAH MASA DEPAN 879
Capaian Pembelajaran Bab 879
Istilah Kunci 879
10.1 Kategori Perangkat Informatika Kimia 880

Pendalaman I — Objek, Makna, dan Tingkat Representasi 882
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 884
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 885
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan887
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 888
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 890
10.2 Open Source, Komersial, dan Strategi Hibrida 891
Pendalaman I — Objek, Makna, dan Tingkat Representasi 893
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 894
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 896
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan897
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 899
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 900
10.3 Python, Notebook, dan RDKit sebagai Lingkungan Belajar..... 901
Pendalaman I — Objek, Makna, dan Tingkat Representasi 904
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 905
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 907
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan908
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 910
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 911
10.4 Pipeline Data dan Workflow Otomatis 912
Pendalaman I — Objek, Makna, dan Tingkat Representasi 914
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 916
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 917
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan919
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 920
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 922
10.5 Version Control, Environment, dan Container..... 923
Pendalaman I — Objek, Makna, dan Tingkat Representasi 925
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 927
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 928
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan930
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 931
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian 932
10.6 Pengujian, Validasi Perangkat, dan Audit Trail 934
Pendalaman I — Objek, Makna, dan Tingkat Representasi 936
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 937

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 939
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan940
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 942
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 943
10.7 Kecerdasan Buatan Generatif dan Model Fondasi Kimia 944
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 946
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 948
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 950
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan951
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 953
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 954
10.8 Reproducible Research, Open Science, dan Publikasi Data 955
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 957
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 959
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 960
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan962
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 963
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 964
10.9 Peta Jalan Kompetensi dan Profesi Informatika Kimia.....966
Pendalaman I — Objek, Makna, dan Tingkat Representasi..... 968
Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja 969
Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi 971
Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan972
Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel 974
Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian..... 975
Praktikum 10 — Paket Riset Reprodusibel.....977
Tujuan..... 977
Bahan dan Perangkat 977
Langkah Kerja 977
Luaran Wajib 977
Pertanyaan Analisis 978
Ringkasan Bab.....978
Latihan dan Tugas Individu.....978
Refleksi Bab 979
LAMPIRAN A. PANDUAN INSTALASI DAN LINGKUNGAN KOMPUTASI 980
LAMPIRAN B. RINGKASAN FORMAT DAN IDENTIFIER KIMIA 981
LAMPIRAN C. KARTU METODE DAN METRIK.....982
LAMPIRAN D. SEPULUH PROYEK BUKU MINI MAHASISWA.....983
D.1 Bahasa Digital Molekul: Dari Graf Kimia ke Identitas Struktur.....983

D.2 Dunia Tiga Dimensi Molekul: Konformer, Farmakofor, dan Visualisasi983
D.3 Arsitektur Basis Data Kimia yang Bersih dan Terhubung 983
D.4 Atlas Ruang Kimia: Deskriptor, Fingerprint, dan Reduksi Dimensi...984
D.5 Mengukur Kedekatan Molekul: Kemiripan, Activity Cliff, dan Keragaman 984
D.6 QSAR yang Dapat Dipercaya: Dari Endpoint ke Model Card.....984
D.7 Penapisan Virtual yang Bertanggung Jawab: Dari HTS ke Kandidat Eksperimen.....984
D.8 Merancang Pustaka Senyawa: Enumerasi, Pareto, dan Retrosintesis .985
D.9 Laboratorium Kimia Digital: ELN, LIMS, Otomasi, dan Siklus DMTA985
D.10 Masa Depan Informatika Kimia Indonesia: Open Science, AI, dan Pendidikan.....985
LAMPIRAN E. SEPULUH TUGAS KELOMPOK.....987
E.1 Audit kualitas 200 rekaman kimia 987
E.2 Kamus konversi format struktur 987
E.3 Eksperimen konformer terkontrol.....987
E.4 Prototipe basis data molekul-assay 987
E.5 Peta ruang kimia kelas senyawa 988
E.6 Benchmark fingerprint dan metrik.....988
E.7 Baseline QSAR dan validasi ketat 988
E.8 Triase virtual multi-tahap.....989
E.9 Pustaka kombinatorial 1.000 produk 989
E.10 Paket reproduksi penelitian.....989
LAMPIRAN F. RUBRIK PENILAIAN PROYEK DAN TUGAS 990
LAMPIRAN G. TEMPLATE LAPORAN DAN DATA CARD 991
LAMPIRAN H. CHECKLIST QUALITY CONTROL 992
LAMPIRAN I. GLOSARIUM A–Z 993
LAMPIRAN J. DAFTAR PUSTAKA.....999
INDEKS TOPIK TERPILIH 1005

DAFTAR GAMBAR

- Gambar 1.1 Transformasi data menjadi keputusan kimia.
- Gambar 1.2 Ekosistem informatika kimia.
- Gambar 2.1 Contoh representasi 2D beberapa molekul.
- Gambar 2.2 Alur pencarian struktur.
- Gambar 3.1 Permukaan energi torsi.
- Gambar 3.2 Model farmakofor.
- Gambar 4.1 Skema basis data relasional.
- Gambar 4.2 Alur kurasi data.
- Gambar 5.1 Proyeksi ruang kimia.
- Gambar 5.2 Fingerprint biner.
- Gambar 6.1 Matriks kemiripan.
- Gambar 6.2 Klastering dan wakil.
- Gambar 7.1 Siklus QSAR/QSPR.
- Gambar 7.2 Validasi eksperimen-prediksi.
- Gambar 8.1 Alur penapisan virtual.
- Gambar 8.2 Profil multi-parameter.
- Gambar 9.1 Enumerasi pustaka.
- Gambar 9.2 Front Pareto.
- Gambar 10.1 Ekosistem perangkat.
- Gambar 10.2 Siklus kerja reprodusibel.

BAB 1. LANDASAN INFORMATIKA KIMIA DAN EKOSISTEM DATA

Orientasi Bab

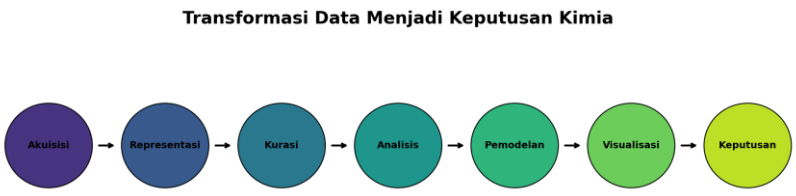
Bab ini menyusun fondasi konseptual dan praktis mengenai landasan informatika kimia. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti landasan informatika kimia dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

audit, basis data struktur, bias, bioaktivitas, data biologis, data kimia, FAIR, high-throughput, identifier, informasi, integritas, interdisiplin, interoperabilitas, kemometrika, keputusan, lisensi, literasi digital, metadata, metode, model fisik, pemrograman, pengetahuan, pertanyaan riset, provenance, QSAR, reaksi, representasi, reproducibility, skala populasi, spektrum, statistika, struktur, transparansi, validasi, visualisasi.



Gambar 1.1. Transformasi data mentah menjadi informasi, model, dan keputusan kimia.

1.1 Definisi, Ruang Lingkup, dan Identitas Disiplin

Informatika kimia sebagai integrasi ilmu kimia, ilmu informasi, statistika, dan komputasi untuk mengubah data molekul menjadi pengetahuan yang dapat dipakai dalam keputusan ilmiah. Pokok bahasan ini menempatkan data kimia, informasi, pengetahuan, dan keputusan sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah data kimia, informasi, pengetahuan, dan keputusan tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah penentuan senyawa prioritas dari ribuan kandidat. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah penyempitan informatika kimia hanya menjadi gambar struktur atau penggunaan perangkat lunak tertentu. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membedakan data, metadata, informasi, model, dan keputusan dalam sebuah studi kimia. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap data kimia, informasi, pengetahuan, dan keputusan harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
data kimia	Peran data kimia pada konteks Definisi, Ruang Lingkup, dan Identitas Disiplin.
informasi	Peran informasi pada konteks Definisi, Ruang Lingkup, dan Identitas Disiplin.
pengetahuan	Peran pengetahuan pada konteks Definisi, Ruang Lingkup, dan Identitas Disiplin.
keputusan	Peran keputusan pada konteks Definisi, Ruang Lingkup, dan Identitas Disiplin.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus penentuan senyawa prioritas dari ribuan kandidat?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 1.1 definisi, ruang lingkup, dan identitas disiplin perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti data kimia dan hubungan operasionalnya dengan informasi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu informatika kimia sebagai integrasi ilmu kimia, ilmu informasi, statistika, dan komputasi untuk mengubah data molekul menjadi pengetahuan yang dapat dipakai dalam keputusan ilmiah, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan data kimia mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar pengetahuan. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah informasi, pengetahuan, dan keputusan harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus penentuan senyawa prioritas dari ribuan kandidat, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu,

misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan data kimia–data kimia mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan data kimia, informasi, pengetahuan, keputusan sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 1.1 definisi, ruang lingkup, dan identitas disiplin perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk penentuan senyawa prioritas dari ribuan kandidat, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi data kimia, atau konversi informasi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan pengetahuan sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa penyempitan informatika kimia hanya menjadi gambar struktur atau penggunaan perangkat lunak tertentu; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 1.1 definisi, ruang lingkup, dan identitas disiplin dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk penentuan senyawa prioritas dari ribuan kandidat, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk

input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.1 definisi, ruang lingkup, dan identitas disiplin, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk penentuan senyawa prioritas dari ribuan kandidat, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai data kimia–data kimia, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus penentuan senyawa prioritas dari ribuan kandidat dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas membedakan data, metadata, informasi, model, dan keputusan dalam sebuah studi kimia sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.1 definisi, ruang lingkup, dan identitas disiplin sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang data kimia akan memengaruhi cara informasi dihitung, cara pengetahuan dibandingkan, dan cara keputusan dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada penentuan senyawa prioritas dari ribuan kandidat', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi,

sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi data kimia, informasi, pengetahuan, keputusan dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
data kimia	Bagaimana data kimia didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.2 Sejarah Perkembangan dan Akar Keilmuan

Perkembangan bidang ini berakar pada penyimpanan struktur, pencarian substruktur, qsar, kemometrika, kimia kombinatorial, dan ledakan data high-throughput. Pokok bahasan ini menempatkan QSAR, basis data struktur, kemometrika, dan high-throughput sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai

hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah QSAR, basis data struktur, kemometrika, dan high-throughput tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah membaca sejarah sebagai daftar perangkat tanpa memahami masalah ilmiah yang mendorong inovasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada

menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun garis waktu hubungan teknologi komputasi dan perubahan cara kerja kimia. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap QSAR, basis data struktur, kemometrika, dan high-throughput harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 1.2 sejarah perkembangan dan akar keilmuan perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti QSAR dan hubungan operasionalnya dengan basis data struktur. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu perkembangan bidang ini berakar pada penyimpanan struktur, pencarian substruktur, QSAR, kemometrika, kimia kombinatorial, dan ledakan data high-throughput, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan QSAR mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar kemometrika. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah basis data struktur, kemometrika, dan high-throughput harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan QSAR–QSAR mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan QSAR, basis data struktur, kemometrika, high-throughput sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 1.2 sejarah perkembangan dan akar keilmuan perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi QSAR, atau konversi basis data struktur—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan kemometrika sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi.

Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa membaca sejarah sebagai daftar perangkat tanpa memahami masalah ilmiah yang mendorong inovasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 1.2 sejarah perkembangan dan akar keilmuan dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi

menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.2 sejarah perkembangan dan akar keilmuan, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai QSAR–QSAR, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas menyusun garis waktu hubungan teknologi komputasi dan perubahan cara kerja kimia sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.2 sejarah perkembangan dan akar keilmuan sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang QSAR akan memengaruhi cara basis data struktur dihitung, cara kemometrika dibandingkan, dan cara high-throughput dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi QSAR, basis data struktur, kemometrika, high-throughput dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
QSAR	Bagaimana QSAR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
high-throughput	Bagaimana high-throughput memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.3 Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data

Batas antardisiplin bersifat permeabel: kimia komputasi menekankan model fisik, bioinformatika menekankan biomolekul dan data biologis, sedangkan informatika kimia menekankan representasi serta pengelolaan populasi molekul dalam skala besar. Pokok bahasan ini menempatkan interdisiplin, model fisik, data biologis, dan skala populasi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah interdisiplin, model fisik, data biologis, dan skala populasi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting

dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah integrasi struktur ligan, target protein, dan hasil uji biologis. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap satu disiplin dapat menggantikan disiplin lain tanpa memperhatikan level representasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah memetakan satu masalah riset ke kontribusi masing-masing disiplin. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap interdisiplin, model fisik, data biologis, dan skala populasi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
interdisiplin	Peran interdisiplin pada konteks Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data.
model fisik	Peran model fisik pada konteks Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data.
data biologis	Peran data biologis pada konteks Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data.
skala populasi	Peran skala populasi pada konteks Hubungan dengan Kimia Komputasi, Bioinformatika, dan Sains Data.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus integrasi struktur ligan, target protein, dan hasil uji biologis?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam analisis yang lebih rinci, unit analisis pada 1.3 hubungan dengan kimia komputasi, bioinformatika, dan sains data perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti interdisiplin dan hubungan operasionalnya dengan model fisik. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif kimia, representasi selalu bersifat selektif. Fokus bagian ini, yaitu batas antardisiplin bersifat permeabel: kimia komputasi menekankan model fisik, bioinformatika menekankan biomolekul dan data biologis, sedangkan informatika kimia menekankan representasi serta pengelolaan populasi molekul dalam skala

besar, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan interdisiplin mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar data biologis. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif komputasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah model fisik, data biologis, dan skala populasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus integrasi struktur ligan, target protein, dan hasil uji biologis, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada tahap verifikasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan interdisiplin–interdisiplin mengubah jumlah entitas unik serta interpretasi hasil.

Untuk pembelajaran berbasis proyek, peta konsep pada bagian ini dapat dibangun dengan menempatkan interdisiplin, model fisik, data biologis, skala populasi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam skenario industri, sumber data untuk 1.3 hubungan dengan kimia komputasi, bioinformatika, dan sains data perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk integrasi struktur ligan, target protein, dan hasil uji biologis, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam riset akademik, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi interdisiplin, atau konversi model fisik—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pertama, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan data biologis sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Selanjutnya, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap satu disiplin dapat menggantikan disiplin lain tanpa memperhatikan level representasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada sisi lain, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam analisis yang lebih rinci, algoritme pada 1.3 hubungan dengan kimia komputasi, bioinformatika, dan sains data dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk integrasi struktur ligan, target protein, dan hasil uji

biologis, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif kimia, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif komputasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada tahap verifikasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Untuk pembelajaran berbasis proyek, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam skenario industri, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.3 hubungan dengan kimia komputasi, bioinformatika, dan sains data, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam riset akademik, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk integrasi struktur ligan, target protein, dan hasil uji biologis, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pertama, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Selanjutnya, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai interdisiplin–interdisiplin, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada sisi lain, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam analisis yang lebih rinci, studi kasus integrasi struktur ligan, target protein, dan hasil uji biologis dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif kimia, aktivitas memetakan satu masalah riset ke kontribusi masing-masing disiplin sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif komputasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada tahap verifikasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Untuk pembelajaran berbasis proyek, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.3 hubungan dengan kimia komputasi, bioinformatika, dan sains data sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang interdisiplin akan memengaruhi cara model fisik dihitung, cara data biologis dibandingkan, dan cara skala populasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada integrasi struktur ligan, target protein, dan hasil uji biologis', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi interdisiplin, model fisik, data biologis, skala populasi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interdisiplin	Bagaimana interdisiplin didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
skala populasi	Bagaimana skala populasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.4 Spektrum Data Kimia dan Siklus Hidup Data

Data kimia mencakup struktur, reaksi, spektrum, sifat fisikokimia, bioaktivitas, toksisitas, dokumen paten, dan catatan eksperimen yang selalu disertai konteks. Pokok bahasan ini menempatkan struktur, reaksi, spektrum, bioaktivitas, dan provenance sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah struktur, reaksi, spektrum, bioaktivitas, dan provenance tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

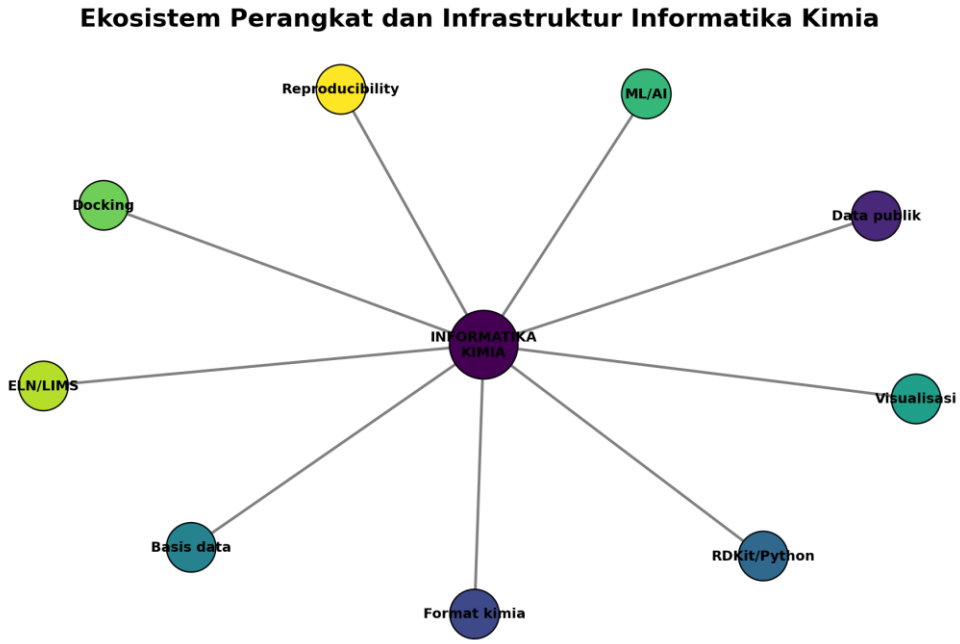
Kesalahan yang sering muncul adalah mengabaikan metadata sehingga angka dari sumber berbeda dianggap sebanding. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat inventaris tipe data dan metadata minimum untuk sebuah eksperimen. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap struktur, reaksi, spektrum, bioaktivitas, dan provenance harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 1.2. Ekosistem disiplin, perangkat, data, dan infrastruktur informatika kimia.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 1.4 spektrum data kimia dan siklus hidup data perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti struktur dan hubungan operasionalnya dengan reaksi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan

versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu data kimia mencakup struktur, reaksi, spektrum, sifat fisikokimia, bioaktivitas, toksisitas, dokumen paten, dan catatan eksperimen yang selalu disertai konteks, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan struktur mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar spektrum. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah reaksi, spektrum, dan bioaktivitas harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan struktur–provenance mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan struktur, reaksi, spektrum, bioaktivitas, provenance sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya,

pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 1.4 spektrum data kimia dan siklus hidup data perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi struktur, atau konversi reaksi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan spektrum sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah

yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengabaikan metadata sehingga angka dari sumber berbeda dianggap sebanding; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 1.4 spektrum data kimia dan siklus hidup data dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas;

transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.4 spektrum data kimia dan siklus hidup data, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai struktur–provenance, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas membuat inventaris tipe data dan metadata minimum untuk sebuah eksperimen sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.4 spektrum data kimia dan siklus hidup data sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang struktur akan memengaruhi cara reaksi dihitung, cara spektrum dibandingkan, dan cara bioaktivitas dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi struktur, reaksi, spektrum, bioaktivitas, provenance dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat

dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
struktur	Bagaimana struktur didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
provenance	Bagaimana provenance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.5 Prinsip FAIR, Provenance, dan Interoperabilitas

Data yang dapat ditemukan, diakses, dipertukarkan, dan digunakan ulang membutuhkan identifier stabil, metadata eksplisit, format terbuka, dan jejak asal-usul yang dapat diaudit. Pokok bahasan ini menempatkan FAIR, identifier, metadata, provenance, dan interoperabilitas sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah FAIR, identifier, metadata, provenance, dan interoperabilitas tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi,

penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah penggunaan InChIKey sebagai kunci silang antarbasis data. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyamakan FAIR dengan data selalu terbuka tanpa memperhatikan lisensi dan kerahasiaan. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menilai sebuah dataset menggunakan daftar periksa FAIR sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap FAIR, identifier, metadata, provenance, dan interoperabilitas harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
FAIR	Peran FAIR pada konteks Prinsip FAIR, Provenance, dan Interoperabilitas.
identifier	Peran identifier pada konteks Prinsip FAIR, Provenance, dan Interoperabilitas.
metadata	Peran metadata pada konteks Prinsip FAIR, Provenance, dan Interoperabilitas.
provenance	Peran provenance pada konteks Prinsip FAIR, Provenance, dan Interoperabilitas.
interoperabilitas	Peran interoperabilitas pada konteks Prinsip FAIR, Provenance, dan Interoperabilitas.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus penggunaan InChIKey sebagai kunci silang antarbasis data?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 1.5 prinsip fair, provenance, dan interoperabilitas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti FAIR dan hubungan operasionalnya dengan identifier. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu data yang dapat ditemukan, diakses, dipertukarkan, dan digunakan ulang membutuhkan identifier stabil, metadata eksplisit, format terbuka, dan jejak asal-usul yang dapat diaudit, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan

FAIR mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar metadata. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah identifier, metadata, dan provenance harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus penggunaan InChIKey sebagai kunci silang antarbasis data, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan FAIR–interoperabilitas mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan FAIR, identifier, metadata, provenance, interoperabilitas sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 1.5 prinsip fair, provenance, dan interoperabilitas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk penggunaan InChIKey sebagai kunci silang antarbasis data, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi FAIR, atau konversi identifier—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan metadata sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier,

struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyamakan FAIR dengan data selalu terbuka tanpa memperhatikan lisensi dan kerahasiaan; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 1.5 prinsip fair, provenance, dan interoperabilitas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk penggunaan InChIKey sebagai kunci silang antarbasis data, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.5 prinsip fair, provenance, dan interoperabilitas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk penggunaan InChIKey sebagai kunci silang antarbasis data, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai

FAIR–interoperabilitas, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus penggunaan InChIKey sebagai kunci silang antarbasis data dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas menilai sebuah dataset menggunakan daftar periksa FAIR sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi,

kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.5 prinsip fair, provenance, dan interoperabilitas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang FAIR akan memengaruhi cara identifier dihitung, cara metadata dibandingkan, dan cara provenance dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada penggunaan InChIKey sebagai kunci silang antarbasis data', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi FAIR, identifier, metadata, provenance, interoperabilitas dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
FAIR	Bagaimana FAIR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interoperabilitas	Bagaimana interoperabilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.6 Etika, Lisensi, Privasi, dan Integritas Ilmiah

Penggunaan data dan algoritme harus menghormati hak cipta, lisensi, kerahasiaan, bias data, keterlacakan, serta pembagian tanggung jawab antara manusia dan sistem otomatis. Pokok bahasan ini menempatkan lisensi, bias, integritas, audit, dan transparansi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah lisensi, bias, integritas, audit, dan transparansi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah model toksisitas yang bias karena ruang kimia latih terlalu sempit. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat

kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyajikan keluaran model sebagai fakta tanpa ketidakpastian dan domain aplikasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis pernyataan penggunaan data dan model yang transparan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap lisensi, bias, integritas, audit, dan transparansi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus model toksisitas yang bias karena ruang kimia latih terlalu sempit?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Selanjutnya, unit analisis pada 1.6 etika, lisensi, privasi, dan integritas ilmiah perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti lisensi dan hubungan operasionalnya dengan bias. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada sisi lain, representasi selalu bersifat selektif. Fokus bagian ini, yaitu penggunaan data dan algoritme harus menghormati hak cipta, lisensi, kerahasiaan, bias data, keterlacakan, serta pembagian tanggung jawab antara manusia dan sistem otomatis, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan lisensi mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar integritas. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam analisis yang lebih rinci, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah bias, integritas, dan audit harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus model toksisitas yang bias karena ruang kimia latih terlalu sempit, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif kimia, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat

yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan lisensi–transparansi mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif komputasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan lisensi, bias, integritas, audit, transparansi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada tahap verifikasi, sumber data untuk 1.6 etika, lisensi, privasi, dan integritas ilmiah perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk model toksisitas yang bias karena ruang kimia latih terlalu sempit, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Untuk pembelajaran berbasis proyek, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi lisensi, atau konversi bias—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam skenario industri, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan integritas sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam riset akademik, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyajikan keluaran model sebagai fakta tanpa ketidakpastian dan domain aplikasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pertama, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Selanjutnya, algoritme pada 1.6 etika, lisensi, privasi, dan integritas ilmiah dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk model toksisitas yang bias karena ruang kimia latih terlalu sempit, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada sisi lain, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam analisis yang lebih rinci, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif kimia, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk

input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif komputasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada tahap verifikasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.6 etika, lisensi, privasi, dan integritas ilmiah, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Untuk pembelajaran berbasis proyek, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk model toksisitas yang bias karena ruang kimia latih terlalu sempit, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam skenario industri, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam riset akademik, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai lisensi–transparansi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pertama, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Selanjutnya, studi kasus model toksisitas yang bias karena ruang kimia latih terlalu sempit dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada sisi lain, aktivitas menulis pernyataan penggunaan data dan model yang transparan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam analisis yang lebih rinci, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif kimia, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat

keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif komputasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.6 etika, lisensi, privasi, dan integritas ilmiah sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang lisensi akan memengaruhi cara bias dihitung, cara integritas dibandingkan, dan cara audit dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada model toksisitas yang bias karena ruang kimia latih terlalu sempit', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap

memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi lisensi, bias, integritas, audit, transparansi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
lisensi	Bagaimana lisensi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transparansi	Bagaimana transparansi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.7 Kompetensi Inti Mahasiswa Kimia S1

Kompetensi dasar meliputi literasi struktur digital, pengelolaan data, pemrograman dasar, statistika, visualisasi, evaluasi model, dan komunikasi hasil yang dapat direproduksi. Pokok bahasan ini menempatkan literasi digital, pemrograman, statistika, visualisasi, dan reproducibility sebagai komponen yang harus dipahami

secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam kerangka pengambilan keputusan, istilah literasi digital, pemrograman, statistika, visualisasi, dan reproducibility tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengejar sintaks pemrograman tanpa memahami pertanyaan kimia. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya

mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun peta kompetensi pribadi dan rencana belajar satu semester. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap literasi digital, pemrograman, statistika, visualisasi, dan reproducibility harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
literasi digital	Peran literasi digital pada konteks Kompetensi Inti Mahasiswa Kimia S1.
pemrograman	Peran pemrograman pada konteks Kompetensi Inti Mahasiswa Kimia S1.
statistika	Peran statistika pada konteks Kompetensi Inti Mahasiswa Kimia S1.
visualisasi	Peran visualisasi pada konteks Kompetensi Inti Mahasiswa Kimia S1.
reproducibility	Peran reproducibility pada konteks Kompetensi Inti Mahasiswa Kimia S1.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 1.7 kompetensi inti mahasiswa kimia s1 perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti literasi digital dan hubungan operasionalnya dengan pemrograman. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu kompetensi dasar meliputi literasi struktur digital, pengelolaan data, pemrograman dasar, statistika, visualisasi, evaluasi model, dan komunikasi hasil yang dapat direproduksi, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan literasi digital mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar statistika. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah pemrograman, statistika, dan visualisasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan literasi digital–reproducibility mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan literasi digital, pemrograman, statistika, visualisasi, reproducibility sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 1.7 kompetensi inti mahasiswa kimia s1 perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi literasi digital, atau konversi pemrograman—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah

sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan statistika sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengejar sintaks pemrograman tanpa memahami pertanyaan kimia; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 1.7 kompetensi inti mahasiswa kimia s1 dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode.

Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.7 kompetensi inti mahasiswa kimia s1, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi, mahasiswa perlu menjelaskan

unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai literasi digital–reproducibility, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas menyusun peta kompetensi pribadi dan rencana belajar satu semester sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang

dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.7 kompetensi inti mahasiswa kimia s1 sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang literasi digital akan memengaruhi cara pemrograman dihitung, cara statistika dibandingkan, dan cara visualisasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi literasi digital, pemrograman, statistika, visualisasi, reproducibility dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
literasi digital	Bagaimana literasi digital didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

1.8 Kerangka Pemecahan Masalah Informatika Kimia

Setiap proyek perlu dirumuskan sebagai rangkaian pertanyaan: apa objek kimianya, representasi apa yang sahih, data apa yang tersedia, metode apa yang cocok, bagaimana validasi dilakukan, dan keputusan apa yang akan dibuat. Pokok bahasan ini menempatkan pertanyaan riset, representasi, metode, validasi, dan keputusan sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya,

melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah pertanyaan riset, representasi, metode, validasi, dan keputusan tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memulai dari algoritme favorit lalu mencari masalah yang dapat dipaksakan. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada

menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis kanvas masalah satu halaman untuk topik riset pilihan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap pertanyaan riset, representasi, metode, validasi, dan keputusan harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 1.8 kerangka pemecahan masalah informatika kimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti pertanyaan riset dan hubungan operasionalnya dengan representasi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu setiap proyek perlu dirumuskan sebagai rangkaian pertanyaan: apa objek kimianya, representasi apa yang sahih, data apa yang tersedia, metode apa yang cocok, bagaimana validasi dilakukan, dan keputusan apa yang akan dibuat, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan pertanyaan riset mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar metode. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah representasi, metode, dan validasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan pertanyaan riset–keputusan mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan pertanyaan riset, representasi, metode, validasi, keputusan sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 1.8 kerangka pemecahan masalah informatika kimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi pertanyaan riset, atau konversi representasi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan metode sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu

dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memulai dari algoritme favorit lalu mencari masalah yang dapat dipaksakan; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 1.8 kerangka pemecahan masalah informatika kimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi

menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 1.8 kerangka pemecahan masalah informatika kimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat

sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai pertanyaan riset–keputusan, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen.

Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas menulis kanvas masalah satu halaman untuk topik riset pilihan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 1.8 kerangka pemecahan masalah informatika kimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang pertanyaan riset akan memengaruhi cara representasi dihitung, cara metode dibandingkan, dan cara validasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi pertanyaan riset, representasi, metode, validasi, keputusan dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat

dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pertanyaan riset	Bagaimana pertanyaan riset didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
keputusan	Bagaimana keputusan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Praktikum 1 — Audit Siklus Hidup Data Kimia

Menganalisis satu dataset senyawa sederhana, memisahkan data, metadata, provenance, dan keputusan; menghasilkan data dictionary dan checklist FAIR.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.

Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Landasan Informatika Kimia harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

Jelaskan dengan contoh hubungan antara data kimia dan keputusan pada konteks penentuan senyawa prioritas dari ribuan kandidat.
Identifikasi minimal tiga risiko jika peneliti penyempitan informatika kimia hanya menjadi gambar struktur atau penggunaan perangkat lunak tertentu.
Jelaskan dengan contoh hubungan antara QSAR dan high-throughput pada konteks peralihan dari arsip kartu dan indeks manual menuju basis data berjuta struktur.

Identifikasi minimal tiga risiko jika peneliti membaca sejarah sebagai daftar perangkat tanpa memahami masalah ilmiah yang mendorong inovasi.
Jelaskan dengan contoh hubungan antara interdisiplin dan skala populasi pada konteks integrasi struktur ligan, target protein, dan hasil uji biologis.
Identifikasi minimal tiga risiko jika peneliti menganggap satu disiplin dapat menggantikan disiplin lain tanpa memperhatikan level representasi.
Jelaskan dengan contoh hubungan antara struktur dan provenance pada konteks rekaman kelarutan yang tidak bermakna tanpa pH, suhu, metode, dan satuan.
Identifikasi minimal tiga risiko jika peneliti mengabaikan metadata sehingga angka dari sumber berbeda dianggap sebanding.
Jelaskan dengan contoh hubungan antara FAIR dan interoperabilitas pada konteks penggunaan InChIKey sebagai kunci silang antarbasis data.
Identifikasi minimal tiga risiko jika peneliti menyamakan FAIR dengan data selalu terbuka tanpa memperhatikan lisensi dan kerahasiaan.
Jelaskan dengan contoh hubungan antara lisensi dan transparansi pada konteks model toksisitas yang bias karena ruang kimia latih terlalu sempit.
Identifikasi minimal tiga risiko jika peneliti menyajikan keluaran model sebagai fakta tanpa ketidakpastian dan domain aplikasi.
Jelaskan dengan contoh hubungan antara literasi digital dan reproducibility pada konteks mahasiswa mengubah data senyawa dari spreadsheet menjadi analisis yang terdokumentasi.
Identifikasi minimal tiga risiko jika peneliti mengejar sintaks pemrograman tanpa memahami pertanyaan kimia.
Jelaskan dengan contoh hubungan antara pertanyaan riset dan keputusan pada konteks pemilihan kandidat inhibitor berdasarkan struktur, aktivitas, dan sifat ADMET.
Identifikasi minimal tiga risiko jika peneliti memulai dari algoritme favorit lalu mencari masalah yang dapat dipaksakan.
Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 2. REPRESENTASI DAN MANIPULASI STRUKTUR MOLEKUL 2D

Orientasi Bab

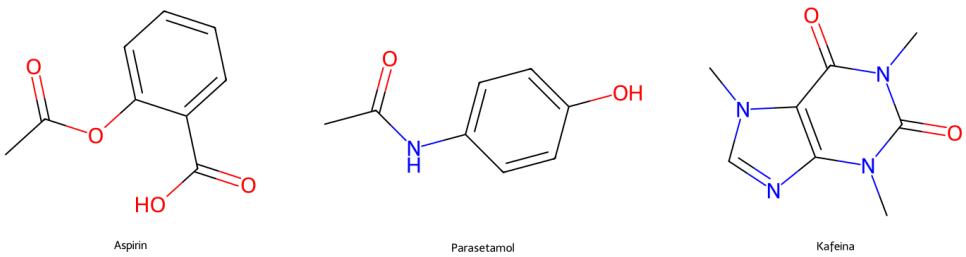
Bab ini menyusun fondasi konseptual dan praktis mengenai representasi molekul 2d. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti representasi molekul 2d dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

ambiguity, aromatisitas, atom mapping, cakupan klaim, connection table, E/Z, garam, InChI, InChIKey, isomorfisme, kandidat, kanonisasi, kiralitas, label atom, Markush, metadata, MOL, orde ikatan, paten, produk, properti, protonasi, pusat reaksi, R-group, R/S, racemat, reaktan, rendemen, screening, SDF, simpul, sisi, SMARTS, SMILES, stereoisomer, subgraf, tautomer, topologi, valensi, verifikasi.



Gambar 2.1. Representasi struktur 2D aspirin, parasetamol, dan kafeina yang dibuat dengan RDKit.

2.1 Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi

Graf molekul menyediakan abstraksi yang memungkinkan struktur kimia disimpan, dibandingkan, dicari, dan dimanipulasi secara algoritmik. Pokok bahasan ini menempatkan simpul, sisi, label atom, orde ikatan, dan topologi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah simpul, sisi, label atom, orde ikatan, dan topologi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap gambar dua dimensi tertentu sebagai satu-satunya representasi graf yang benar. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membangun daftar atom dan ikatan untuk molekul sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap simpul, sisi, label atom, orde ikatan, dan topologi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
simpul	Peran simpul pada konteks Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi.
sisi	Peran sisi pada konteks Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi.
label atom	Peran label atom pada konteks Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi.
orde ikatan	Peran orde ikatan pada konteks Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi.
topologi	Peran topologi pada konteks Graf Molekul: Atom sebagai Simpul dan Ikatan sebagai Sisi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 2.1 graf molekul: atom sebagai simpul dan ikatan sebagai sisi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti simpul dan hubungan operasionalnya dengan sisi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu graf molekul menyediakan abstraksi yang memungkinkan struktur kimia disimpan, dibandingkan, dicari, dan dimanipulasi secara algoritmik, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan simpul mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar label atom. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah sisi, label atom, dan orde ikatan harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi

berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan simpul–topologi mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan simpul, sisi, label atom, orde ikatan, topologi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 2.1 graf molekul: atom sebagai simpul dan ikatan sebagai sisi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas.

Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi simpul, atau konversi sisi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan label atom sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap gambar dua dimensi tertentu sebagai satu-satunya representasi graf yang benar; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 2.1 graf molekul: atom sebagai simpul dan ikatan sebagai sisi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap

akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.1 graf molekul:

atom sebagai simpul dan ikatan sebagai sisi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai simpul–topologi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas membangun daftar atom dan ikatan untuk molekul sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi

seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.1 graf molekul: atom sebagai simpul dan ikatan sebagai sisi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang simpul akan memengaruhi cara sisi dihitung, cara label atom dibandingkan, dan cara orde ikatan dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta

ikatan tunggal, rangkap, dan aromatik', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi simpul, sisi, label atom, orde ikatan, topologi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
simpul	Bagaimana simpul didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
topologi	Bagaimana topologi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.2 Tabel Konektivitas dan Format Berkas

Format mol, sdf, mol2, dan format pertukaran lain menyimpan atom, ikatan, koordinat, properti, dan metadata dengan tingkat kelengkapan berbeda. Pokok bahasan ini menempatkan connection table, MOL, SDF, properti, dan metadata sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah connection table, MOL, SDF, properti, dan metadata tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah kehilangan muatan, stereokimia, atau atom hidrogen saat konversi format. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan informasi yang dipertahankan oleh SMILES dan SDF. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap connection table, MOL, SDF, properti, dan metadata harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 2.2 tabel konektivitas dan format berkas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti connection table dan hubungan operasionalnya dengan MOL. Bila unit analisis berubah di tengah

pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu format MOL, SDF, MOL2, dan format pertukaran lain menyimpan atom, ikatan, koordinat, properti, dan metadata dengan tingkat kelengkapan berbeda, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan connection table mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar SDF. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah MOL, SDF, dan properti harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan connection table–metadata mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan connection table, MOL, SDF, properti, metadata sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman

mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 2.2 tabel konektivitas dan format berkas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi connection table, atau konversi MOL—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan SDF sebelum koreksi struktur dapat mempertahankan kesalahan ke

tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa kehilangan muatan, stereokimia, atau atom hidrogen saat konversi format; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 2.2 tabel konektivitas dan format berkas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan

keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.2 tabel konektivitas dan format berkas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai connection table–metadata, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membandingkan informasi yang dipertahankan oleh SMILES dan SDF sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.2 tabel konektivitas dan format berkas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang connection table akan memengaruhi cara MOL dihitung, cara SDF dibandingkan, dan cara properti dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi connection table, MOL, SDF, properti, metadata dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
connection table	Bagaimana connection table didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
metadata	Bagaimana metadata memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.3 Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey

Notasi linear memungkinkan struktur dikomunikasikan sebagai teks, tetapi tujuan smiles, smarts, dan inchi berbeda sehingga tidak boleh dipertukarkan tanpa memahami semantiknya. Pokok bahasan ini menempatkan SMILES, SMARTS, InChI, InChIKey, dan kanonisasi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam kerangka pengambilan keputusan, istilah SMILES, SMARTS, InChI, InChIKey, dan kanonisasi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan SMILES nonkanonik sebagai identifier permanen. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengonversi beberapa struktur ke notasi linear dan memeriksa stereokimia. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap SMILES, SMARTS, InChI, InChIKey, dan kanonisasi harus selalu dihubungkan dengan bab lain. Representasi

memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
SMILES	Peran SMILES pada konteks Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey.
SMARTS	Peran SMARTS pada konteks Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey.
InChI	Peran InChI pada konteks Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey.
InChIKey	Peran InChIKey pada konteks Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey.
kanonisasi	Peran kanonisasi pada konteks Notasi Linear: SMILES, SMARTS, InChI, dan InChIKey.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 2.3 notasi linear: smiles, smarts, inchi, dan inchikey perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti SMILES dan hubungan operasionalnya dengan SMARTS. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu notasi linear memungkinkan struktur dikomunikasikan sebagai teks, tetapi tujuan SMILES, SMARTS, dan InChI berbeda sehingga tidak boleh dipertukarkan tanpa memahami semantiknya, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan SMILES mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar InChI. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan,

waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah SMARTS, InChI, dan InChIKey harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan SMILES–kanonisasi mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan SMILES, SMARTS, InChI, InChIKey, kanonisasi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 2.3 notasi linear: smiles, smarts, inchi, dan inchikey perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi SMILES, atau konversi SMARTS—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan InChI sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak,

outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan SMILES nonkanonik sebagai identifier permanen; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 2.3 notasi linear: smiles, smarts, inchi, dan inchikey dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.3 notasi linear: smiles, smarts, inchi, dan inchikey, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster.

Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai SMILES–kanonisasi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas mengonversi beberapa struktur ke notasi linear dan memeriksa stereokimia sebaiknya menghasilkan lebih dari laporan naratif. Portofolio

minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.3 notasi linear: smiles, smarts, inchi, dan inchikey sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang SMILES akan memengaruhi cara SMARTS dihitung, cara InChI dibandingkan, dan cara InChIKey dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi SMILES, SMARTS, InChI, InChIKey, kanonisasi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SMILES	Bagaimana SMILES didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kanonisasi	Bagaimana kanonisasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.4 Aromatisitas, Valensi, Muatan, Tautomer, dan Protonasi

Normalisasi kimia harus menangani model aromatisitas, valensi, muatan formal, tautomerisme, keadaan protonasi, dan garam secara konsisten sesuai tujuan analisis. Pokok bahasan ini menempatkan aromatisitas, valensi, tautomer, protonasi, dan garam sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah aromatisitas, valensi, tautomer, protonasi, dan garam tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mencampur bentuk garam dan parent compound sebagai entitas identik dalam model. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

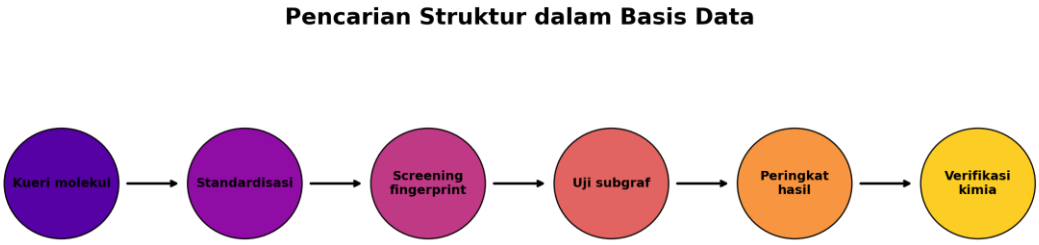
Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah merancang aturan standardisasi untuk sebuah koleksi senyawa. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap aromatisitas, valensi, tautomer, protonasi, dan garam harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia

adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 2.2. Screening fingerprint memperkecil kandidat sebelum verifikasi subgraf.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 2.4 aromatisitas, valensi, muatan, tautomer, dan protonasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti aromatisitas dan hubungan operasionalnya dengan valensi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu normalisasi kimia harus menangani model aromatisitas, valensi, muatan formal, tautomerisme, keadaan protonasi, dan garam secara konsisten sesuai tujuan analisis, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan aromatisitas mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar tautomer. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan

informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah valensi, tautomer, dan protonasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan aromatisitas–garam mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan aromatisitas, valensi, tautomer, protonasi, garam sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 2.4 aromatisitas, valensi, muatan, tautomer, dan protonasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi aromatisitas, atau konversi valensi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan tautomer sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mencampur bentuk garam dan parent compound

sebagai entitas identik dalam model; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 2.4 aromatisitas, valensi, muatan, tautomer, dan protonasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis

sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.4 aromatisitas, valensi, muatan, tautomer, dan protonasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai aromatisitas–garam, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas merancang aturan standardisasi untuk sebuah koleksi senyawa sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.4 aromatisitas, valensi, muatan, tautomer, dan protonasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang aromatisitas akan memengaruhi cara valensi dihitung, cara tautomer dibandingkan,

dan cara protonasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi aromatisitas, valensi, tautomer, protonasi, garam dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
aromatisitas	Bagaimana aromatisitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
garam	Bagaimana garam memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.5 Stereokimia dan Isomerisme dalam Representasi Digital

Isomer r/s, e/z, atropisomer, dan pusat stereogenik lain membutuhkan penandaan eksplisit agar identitas dan aktivitas biologis tidak tertukar. Pokok bahasan ini menempatkan kiralitas, R/S, E/Z, stereoisomer, dan racemat sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah kiralitas, R/S, E/Z, stereoisomer, dan racemat tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah dua enansiomer dengan aktivitas farmakologis berbeda. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini

mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menghapus informasi stereokimia ketika mengekspor SMILES. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah memeriksa apakah suatu format mempertahankan stereokimia. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap kiralitas, R/S, E/Z, stereoisomer, dan racemat harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
kiralitas	Peran kiralitas pada konteks Stereokimia dan Isomerisme dalam Representasi Digital.

Konsep	Makna operasional dalam bagian ini
R/S	Peran R/S pada konteks Stereokimia dan Isomerisme dalam Representasi Digital.
E/Z	Peran E/Z pada konteks Stereokimia dan Isomerisme dalam Representasi Digital.
stereoisomer	Peran stereoisomer pada konteks Stereokimia dan Isomerisme dalam Representasi Digital.
racemat	Peran racemat pada konteks Stereokimia dan Isomerisme dalam Representasi Digital.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus dua enansiomer dengan aktivitas farmakologis berbeda?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 2.5 stereokimia dan isomerisme dalam representasi digital perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti kiralitas dan hubungan operasionalnya dengan R/S. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu isomer R/S, E/Z, atropisomer, dan pusat stereogenik lain membutuhkan penandaan eksplisit agar identitas dan aktivitas biologis tidak tertukar, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan kiralitas mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar E/Z. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah R/S, E/Z, dan stereoisomer harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang

berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus dua enansiomer dengan aktivitas farmakologis berbeda, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan kiralitas–racemat mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan kiralitas, R/S, E/Z, stereoisomer, racemat sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 2.5 stereokimia dan isomerisme dalam representasi digital perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk dua enansiomer dengan aktivitas farmakologis berbeda, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi kiralitas, atau konversi R/S—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan E/Z sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menghapus informasi stereokimia ketika mengekspor SMILES; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat

diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 2.5 stereokimia dan isomerisme dalam representasi digital dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk dua enansiomer dengan aktivitas farmakologis berbeda, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.5 stereokimia dan isomerisme dalam representasi digital, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk dua enansiomer dengan aktivitas farmakologis berbeda, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai kiralitas–racemat, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus dua enansiomer dengan aktivitas farmakologis berbeda dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas memeriksa apakah suatu format mempertahankan stereokimia sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.5 stereokimia dan isomerisme dalam representasi digital sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang kiralitas akan memengaruhi cara R/S dihitung, cara E/Z dibandingkan, dan cara stereoisomer dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada dua enansiomer dengan aktivitas farmakologis berbeda', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi kiralitas, R/S, E/Z, stereoisomer, racemat dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
kiralitas	Bagaimana kiralitas didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
racemat	Bagaimana racemat memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.6 Isomorfisme Graf dan Pencarian Substruktur

Pencarian struktur tepat, substruktur, dan superstruktur merupakan masalah pencocokan graf yang biasanya dipercepat dengan screening fingerprint sebelum algoritme subgraf dijalankan. Pokok bahasan ini menempatkan isomorfisme, subgraf, screening, kandidat, dan verifikasi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam kerangka pengambilan keputusan, istilah isomorfisme, subgraf, screening, kandidat, dan verifikasi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mencari semua senyawa yang memiliki cincin anilida dalam basis data. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap fingerprint match sebagai bukti final tanpa verifikasi subgraf. Kekeliruan tersebut dapat menghasilkan analisis yang

tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat kueri substruktur sederhana dan menafsirkan hasil. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap isomorfisme, subgraf, screening, kandidat, dan verifikasi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mencari semua senyawa yang memiliki cincin anilida dalam basis data?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam analisis yang lebih rinci, unit analisis pada 2.6 isomorfisme graf dan pencarian substruktur perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti isomorfisme dan hubungan operasionalnya dengan subgraf. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif kimia, representasi selalu bersifat selektif. Fokus bagian ini, yaitu pencarian struktur tepat, substruktur, dan superstruktur merupakan masalah pencocokan graf yang biasanya dipercepat dengan screening fingerprint sebelum algoritme subgraf dijalankan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan isomorfisme mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar screening. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif komputasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah subgraf, screening, dan kandidat harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mencari semua senyawa yang memiliki cincin anilida dalam basis data, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada tahap verifikasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan isomorfisme–verifikasi mengubah jumlah entitas unik serta interpretasi hasil.

Untuk pembelajaran berbasis proyek, peta konsep pada bagian ini dapat dibangun dengan menempatkan isomorfisme, subgraf, screening, kandidat, verifikasi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam skenario industri, sumber data untuk 2.6 isomorfisme graf dan pencarian substruktur perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mencari semua senyawa yang memiliki cincin anilida dalam basis data, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam riset akademik, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi isomorfisme, atau konversi subgraf—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada

kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pertama, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan screening sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Selanjutnya, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap fingerprint match sebagai bukti final tanpa verifikasi subgraf; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada sisi lain, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam analisis yang lebih rinci, algoritme pada 2.6 isomorfisme graf dan pencarian substruktur dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mencari semua senyawa yang memiliki cincin anilida dalam basis data, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif kimia, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif komputasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada tahap verifikasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Untuk pembelajaran berbasis proyek, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam skenario industri, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.6 isomorfisme graf dan pencarian substruktur, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam riset akademik, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mencari semua senyawa yang memiliki cincin anilida dalam basis data, mahasiswa perlu menjelaskan unit independensi:

apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pertama, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Selanjutnya, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai isomorfisme–verifikasi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada sisi lain, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam analisis yang lebih rinci, studi kasus mencari semua senyawa yang memiliki cincin anilida dalam basis data dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif kimia, aktivitas membuat kueri substruktur sederhana dan menafsirkan hasil sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif komputasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada tahap verifikasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Untuk pembelajaran berbasis proyek, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.6 isomorfisme graf dan pencarian substruktur sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang isomorfisme akan memengaruhi cara subgraf dihitung, cara screening dibandingkan, dan cara kandidat dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mencari semua senyawa yang memiliki cincin anilida dalam basis data', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi isomorfisme, subgraf, screening, kandidat, verifikasi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas,

membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
isomorfisme	Bagaimana isomorfisme didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
verifikasi	Bagaimana verifikasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.7 Basis Data Reaksi dan Representasi Transformasi

Reaksi perlu merepresentasikan reaktan, produk, agen, atom mapping, pusat reaksi, kondisi, rendemen, dan sumber agar dapat ditelusuri dan dimodelkan. Pokok bahasan ini menempatkan reaktan, produk, atom mapping, pusat reaksi, dan rendemen sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah reaktan, produk, atom mapping, pusat reaksi, dan rendemen tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis.

Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mencampur reagen dan pelarut dengan reaktan utama tanpa peran yang jelas. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun skema data minimum untuk satu reaksi organik. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap reaktan, produk, atom mapping, pusat reaksi, dan rendemen harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
reaktan	Peran reaktan pada konteks Basis Data Reaksi dan Representasi Transformasi.
produk	Peran produk pada konteks Basis Data Reaksi dan Representasi Transformasi.
atom mapping	Peran atom mapping pada konteks Basis Data Reaksi dan Representasi Transformasi.
pusat reaksi	Peran pusat reaksi pada konteks Basis Data Reaksi dan Representasi Transformasi.
rendemen	Peran rendemen pada konteks Basis Data Reaksi dan Representasi Transformasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 2.7 basis data reaksi dan representasi transformasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti reaktan dan hubungan operasionalnya dengan produk. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan

versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu reaksi perlu merepresentasikan reaktan, produk, agen, atom mapping, pusat reaksi, kondisi, rendemen, dan sumber agar dapat ditelusuri dan dimodelkan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan reaktan mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar atom mapping. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah produk, atom mapping, dan pusat reaksi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan reaktan–rendemen mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan reaktan, produk, atom mapping, pusat reaksi, rendemen sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya,

pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 2.7 basis data reaksi dan representasi transformasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi reaktan, atau konversi produk—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan atom mapping sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai

objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mencampur reagen dan pelarut dengan reaktan utama tanpa peran yang jelas; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 2.7 basis data reaksi dan representasi transformasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas;

transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.7 basis data reaksi dan representasi transformasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai reaktan–rendemen, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas menyusun skema data minimum untuk satu reaksi organik sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.7 basis data reaksi dan representasi transformasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang reaktan akan memengaruhi cara produk dihitung, cara atom mapping dibandingkan, dan cara pusat reaksi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi reaktan, produk, atom mapping, pusat reaksi, rendemen dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaktan	Bagaimana reaktan didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rendemen	Bagaimana rendemen memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

2.8 Paten, Struktur Markush, dan Tantangan Ambiguitas

Dokumen paten sering menggunakan struktur generik markush yang mewakili keluarga sangat besar sehingga pencarian memerlukan penalaran struktur dan bahasa hukum. Pokok bahasan ini menempatkan Markush, R-group, cakupan klaim, ambiguity, dan paten sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah Markush, R-group, cakupan klaim, ambiguity, dan paten tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap satu gambar contoh sebagai seluruh cakupan klaim paten. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengidentifikasi elemen variabel dalam skema Markush sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Markush, R-group, cakupan klaim, ambiguity, dan paten harus selalu dihubungkan dengan bab lain. Representasi

memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 2.8 paten, struktur markush, dan tantangan ambiguitas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Markush dan hubungan operasionalnya dengan R-group. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu dokumen paten sering menggunakan struktur generik Markush yang mewakili keluarga sangat besar sehingga pencarian memerlukan penalaran struktur dan bahasa hukum, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Markush mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar cakupan klaim. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah R-group, cakupan klaim, dan ambiguity harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda.

Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Markush–paten mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan Markush, R-group, cakupan klaim, ambiguity, paten sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 2.8 paten, struktur markush, dan tantangan ambiguitas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari

agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Markush, atau konversi R-group—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan cakupan klaim sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap satu gambar contoh sebagai seluruh cakupan klaim paten; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 2.8 paten, struktur markush, dan tantangan ambiguitas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap

akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 2.8

paten, struktur markush, dan tantangan ambiguitas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Markush–paten, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas mengidentifikasi elemen variabel dalam skema Markush sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 2.8 paten, struktur markush, dan tantangan ambiguitas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Markush akan memengaruhi cara R-group dihitung, cara cakupan klaim dibandingkan, dan cara ambiguity dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen', melainkan membandingkan dua kebijakan representasi atau dua

strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Markush, R-group, cakupan klaim, ambiguity, paten dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Markush	Bagaimana Markush didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
paten	Bagaimana paten memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$G_1 \cong G_2$ jika terdapat pemetaan satu-ke-satu simpul yang mempertahankan keterhubungan dan label kimia.

Praktikum 2 — SMILES, SMARTS, dan Pencarian Substruktur

Menggunakan RDKit untuk membaca SMILES, memvalidasi struktur, membuat SMARTS, melakukan pencarian substruktur, dan memeriksa stereokimia.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from rdkit import Chem

smiles = "CC(=O)Oc1ccccc1C(=O)O"
mol = Chem.MolFromSmiles(smiles)
assert mol is not None
print(Chem.MolToSmiles(mol, canonical=True))

query = Chem.MolFromSmarts("C(=O)O")
print(mol.HasSubstructMatch(query))
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.

- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Representasi Molekul 2D harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara simpul dan topologi pada konteks aspirin sebagai graf dengan atom karbon dan oksigen berlabel serta ikatan tunggal, rangkap, dan aromatik.
2. Identifikasi minimal tiga risiko jika peneliti menganggap gambar dua dimensi tertentu sebagai satu-satunya representasi graf yang benar.
3. Jelaskan dengan contoh hubungan antara connection table dan metadata pada konteks SDF berisi banyak rekaman dengan blok struktur dan pasangan nama-nilai.
4. Identifikasi minimal tiga risiko jika peneliti kehilangan muatan, stereokimia, atau atom hidrogen saat konversi format.
5. Jelaskan dengan contoh hubungan antara SMILES dan kanonisasi pada konteks SMILES untuk aspirin, SMARTS untuk gugus ester, dan InChIKey untuk pencocokan identitas.
6. Identifikasi minimal tiga risiko jika peneliti menggunakan SMILES nonkanonik sebagai identifier permanen.

7. Jelaskan dengan contoh hubungan antara aromatisitas dan garam pada konteks perbedaan bentuk netral dan terprotonasi obat basa pada pH fisiologis.
8. Identifikasi minimal tiga risiko jika peneliti mencampur bentuk garam dan parent compound sebagai entitas identik dalam model.
9. Jelaskan dengan contoh hubungan antara kiralitas dan racemat pada konteks dua enansiomer dengan aktivitas farmakologis berbeda.
10. Identifikasi minimal tiga risiko jika peneliti menghapus informasi stereokimia ketika mengekspor SMILES.
11. Jelaskan dengan contoh hubungan antara isomorfisme dan verifikasi pada konteks mencari semua senyawa yang memiliki cincin anilida dalam basis data.
12. Identifikasi minimal tiga risiko jika peneliti menganggap fingerprint match sebagai bukti final tanpa verifikasi subgraf.
13. Jelaskan dengan contoh hubungan antara reaktan dan rendemen pada konteks reaksi esterifikasi sebagai transformasi ikatan dengan atom yang dipetakan.
14. Identifikasi minimal tiga risiko jika peneliti mencampur reagen dan pelarut dengan reaktan utama tanpa peran yang jelas.
15. Jelaskan dengan contoh hubungan antara Markush dan paten pada konteks scaffold dengan beberapa posisi R yang masing-masing memiliki daftar substituen.
16. Identifikasi minimal tiga risiko jika peneliti menganggap satu gambar contoh sebagai seluruh cakupan klaim paten.
17. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
18. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
19. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
20. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 3. STRUKTUR 3D, KONFORMASI, FARMAKOFOR, DAN VISUALISASI

Orientasi Bab

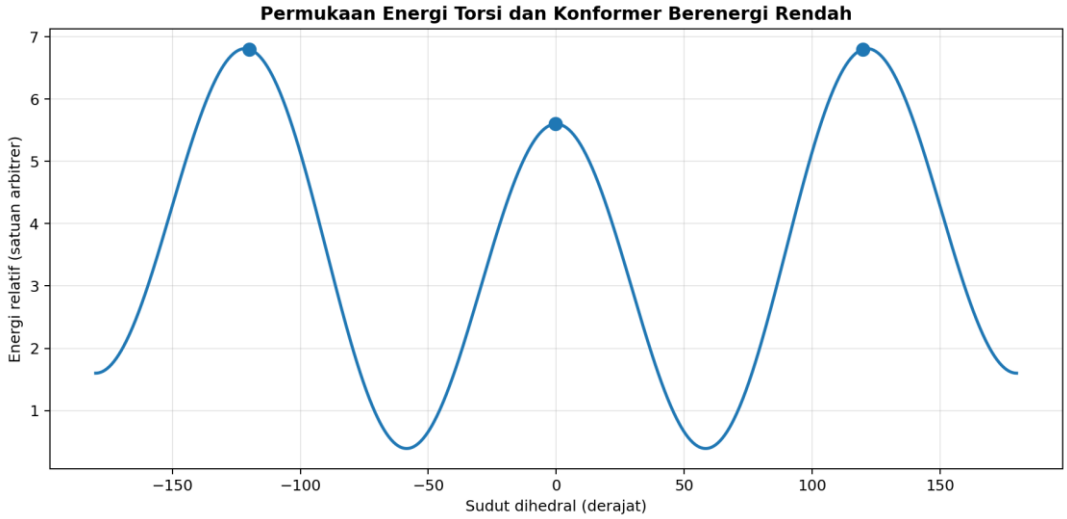
Bab ini menyusun fondasi konseptual dan praktis mengenai struktur molekul 3d. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti struktur molekul 3d dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

acak, akseptor, alignment, annotasi, aromatik, barrier, basis data konformer, dihedral, distance geometry, donor, ensemble, fleksibilitas, hidrofobik, indeks 3D, Kartesian, kendala jarak, ketidakpastian, konformer, koordinat, korespondensi atom, kristalografi, kunci jarak, medan, minimum energi, Monte Carlo, NMR, panjang ikatan, permukaan, rendering, representasi stik, resolusi, RMSD, ruang konformasi, sensitivitas, shape overlap, sistematik, sudut, transformasi, warna atom.



Gambar 3.1. Ilustrasi permukaan energi torsi dan beberapa minimum konformasi.

3.1 Koordinat 3D dan Sumber Struktur Eksperimental

Struktur tiga dimensi berasal dari kristalografi, nmr, cryo-em, atau pemodelan, dan setiap sumber memiliki resolusi, ketidakpastian, serta konteks fisik yang berbeda. Pokok bahasan ini menempatkan koordinat, kristalografi, NMR, resolusi, dan ketidakpastian sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah koordinat, kristalografi, NMR, resolusi, dan ketidakpastian tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah konformasi ligan dalam kristal protein dibandingkan konformasi larutan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap satu struktur kristal sebagai satu-satunya konformasi biologis. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mencatat sumber, resolusi, dan kondisi struktur 3D. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap koordinat, kristalografi, NMR, resolusi, dan ketidakpastian harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku:

informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
koordinat	Peran koordinat pada konteks Koordinat 3D dan Sumber Struktur Eksperimental.
kristalografi	Peran kristalografi pada konteks Koordinat 3D dan Sumber Struktur Eksperimental.
NMR	Peran NMR pada konteks Koordinat 3D dan Sumber Struktur Eksperimental.
resolusi	Peran resolusi pada konteks Koordinat 3D dan Sumber Struktur Eksperimental.
ketidakpastian	Peran ketidakpastian pada konteks Koordinat 3D dan Sumber Struktur Eksperimental.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus konformasi ligan dalam kristal protein dibandingkan konformasi larutan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam analisis yang lebih rinci, unit analisis pada 3.1 koordinat 3d dan sumber struktur eksperimental perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti koordinat dan hubungan operasionalnya dengan kristalografi. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif kimia, representasi selalu bersifat selektif. Fokus bagian ini, yaitu struktur tiga dimensi berasal dari kristalografi, NMR, cryo-EM, atau pemodelan, dan setiap sumber memiliki resolusi, ketidakpastian, serta konteks fisik yang berbeda, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan koordinat mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar NMR. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang

dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif komputasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah kristalografi, NMR, dan resolusi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus konformasi ligan dalam kristal protein dibandingkan konformasi larutan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada tahap verifikasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan koordinat–ketidakpastian mengubah jumlah entitas unik serta interpretasi hasil.

Untuk pembelajaran berbasis proyek, peta konsep pada bagian ini dapat dibangun dengan menempatkan koordinat, kristalografi, NMR, resolusi, ketidakpastian sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam skenario industri, sumber data untuk 3.1 koordinat 3d dan sumber struktur eksperimental perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk konformasi ligan dalam kristal protein dibandingkan konformasi larutan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam riset akademik, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi koordinat, atau konversi kristalografi—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pertama, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan NMR sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Selanjutnya, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap satu struktur kristal sebagai satu-

satunya konformasi biologis; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada sisi lain, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam analisis yang lebih rinci, algoritme pada 3.1 koordinat 3d dan sumber struktur eksperimental dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk konformasi ligan dalam kristal protein dibandingkan konformasi larutan, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif kimia, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis

sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif komputasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada tahap verifikasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Untuk pembelajaran berbasis proyek, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam skenario industri, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.1 koordinat 3d dan sumber struktur eksperimental, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam riset akademik, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk konformasi ligan dalam kristal protein dibandingkan konformasi larutan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pertama, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Selanjutnya, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai koordinat–ketidakpastian, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada sisi lain, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman

eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam analisis yang lebih rinci, studi kasus konformasi ligan dalam kristal protein dibandingkan konformasi larutan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif kimia, aktivitas mencatat sumber, resolusi, dan kondisi struktur 3D sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif komputasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada tahap verifikasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Untuk pembelajaran berbasis proyek, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.1 koordinat 3d dan sumber struktur eksperimental sebagai mata rantai dalam sistem yang lebih besar. Keputusan

tentang koordinat akan memengaruhi cara kristalografi dihitung, cara NMR dibandingkan, dan cara resolusi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada konformasi ligan dalam kristal protein dibandingkan konformasi larutan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi koordinat, kristalografi, NMR, resolusi, ketidakpastian dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
koordinat	Bagaimana koordinat didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ketidakpastian	Bagaimana ketidakpastian memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.2 Geometri Molekul, Koordinat Internal, dan Orientasi

Geometri dapat dinyatakan dengan koordinat kartesian atau koordinat internal berupa panjang ikatan, sudut, dan dihedral yang memengaruhi cara transformasi dilakukan. Pokok bahasan ini menempatkan Kartesian, panjang ikatan, sudut, dihedral, dan transformasi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah Kartesian, panjang ikatan, sudut, dihedral, dan transformasi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini

mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah membandingkan struktur tanpa penyelarasan atau invariansi rotasi-translasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengukur parameter geometri dari satu molekul. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Kartesian, panjang ikatan, sudut, dihedral, dan transformasi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Selanjutnya, unit analisis pada 3.2 geometri molekul, koordinat internal, dan orientasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Kartesian dan hubungan operasionalnya dengan panjang ikatan. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada sisi lain, representasi selalu bersifat selektif. Fokus bagian ini, yaitu geometri dapat dinyatakan dengan koordinat Kartesian atau koordinat internal berupa panjang ikatan, sudut, dan dihedral yang memengaruhi cara transformasi dilakukan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Kartesian mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar sudut. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam analisis yang lebih rinci, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah panjang ikatan, sudut, dan dihedral harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif kimia, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat

yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Kartesian–transformasi mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif komputasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan Kartesian, panjang ikatan, sudut, dihedral, transformasi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada tahap verifikasi, sumber data untuk 3.2 geometri molekul, koordinat internal, dan orientasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Untuk pembelajaran berbasis proyek, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Kartesian, atau konversi panjang ikatan—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam skenario industri, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan sudut sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam riset akademik, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa membandingkan struktur tanpa penyelarasan atau invariansi rotasi-translasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pertama, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Selanjutnya, algoritme pada 3.2 geometri molekul, koordinat internal, dan orientasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada sisi lain, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam analisis yang lebih rinci, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif kimia, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk

input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif komputasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada tahap verifikasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.2 geometri molekul, koordinat internal, dan orientasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Untuk pembelajaran berbasis proyek, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam skenario industri, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam riset akademik, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Kartesian–transformasi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pertama, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Selanjutnya, studi kasus rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada sisi lain, aktivitas mengukur parameter geometri dari satu molekul sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam analisis yang lebih rinci, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif kimia, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat

keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif komputasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.2 geometri molekul, koordinat internal, dan orientasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Kartesian akan memengaruhi cara panjang ikatan dihitung, cara sudut dibandingkan, dan cara dihedral dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan.

Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Kartesian, panjang ikatan, sudut, dihedral, transformasi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kartesian	Bagaimana Kartesian didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
transformasi	Bagaimana transformasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.3 Energi Potensial dan Ruang Konformasi

Molekul fleksibel memiliki banyak minimum lokal pada permukaan energi sehingga pencarian konformer harus menyeimbangkan cakupan, biaya komputasi, dan relevansi lingkungan. Pokok bahasan ini menempatkan minimum energi,

barrier, konformer, ruang konformasi, dan fleksibilitas sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah minimum energi, barrier, konformer, ruang konformasi, dan fleksibilitas tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah butana memiliki konformasi anti dan gauche dengan energi berbeda. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih konformer energi terendah di vakum sebagai konformer aktif tanpa bukti. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh,

menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah memetakan energi torsi untuk satu ikatan rotasi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap minimum energi, barrier, konformer, ruang konformasi, dan fleksibilitas harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
minimum energi	Peran minimum energi pada konteks Energi Potensial dan Ruang Konformasi.
barrier	Peran barrier pada konteks Energi Potensial dan Ruang Konformasi.
konformer	Peran konformer pada konteks Energi Potensial dan Ruang Konformasi.
ruang konformasi	Peran ruang konformasi pada konteks Energi Potensial dan Ruang Konformasi.
fleksibilitas	Peran fleksibilitas pada konteks Energi Potensial dan Ruang Konformasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus butana memiliki konformasi anti dan gauche dengan energi berbeda?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 3.3 energi potensial dan ruang konformasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti minimum energi dan hubungan operasionalnya dengan barrier. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu molekul fleksibel memiliki banyak minimum lokal pada permukaan energi sehingga pencarian konformer harus menyeimbangkan cakupan, biaya komputasi, dan relevansi lingkungan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan minimum energi mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar konformer. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah barrier, konformer, dan ruang konformasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus butana memiliki konformasi anti dan gauche dengan energi berbeda, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat

bagaimana kebijakan minimum energi–fleksibilitas mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan minimum energi, barrier, konformer, ruang konformasi, fleksibilitas sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 3.3 energi potensial dan ruang konformasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk butana memiliki konformasi anti dan gauche dengan energi berbeda, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi minimum energi, atau konversi barrier—diberi nama, alasan, parameter, jumlah

rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan konformer sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih konformer energi terendah di vakum sebagai konformer aktif tanpa bukti; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 3.3 energi potensial dan ruang konformasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk butana memiliki konformasi anti dan gauche dengan energi berbeda, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan

versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.3 energi potensial dan ruang konformasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk butana memiliki konformasi anti dan

gauche dengan energi berbeda, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai minimum energi–fleksibilitas, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus butana memiliki konformasi anti dan gauche dengan energi berbeda dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas memetakan energi torsi untuk satu ikatan rotasi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk

mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.3 energi potensial dan ruang konformasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang minimum energi akan memengaruhi cara barrier dihitung, cara konformer dibandingkan, dan cara ruang konformasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada butana memiliki konformasi anti dan gauche dengan energi berbeda', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi minimum energi, barrier, konformer, ruang konformasi, fleksibilitas dengan data eksperimen,

literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
minimum energi	Bagaimana minimum energi didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fleksibilitas	Bagaimana fleksibilitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.4 Algoritme Pencarian Konformer

Metode sistematik, acak, monte carlo, dinamika molekul, distance geometry, dan algoritme evolusioner memiliki strategi eksplorasi berbeda. Pokok bahasan ini menempatkan sistematik, acak, distance geometry, Monte Carlo, dan ensemble sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu

menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah sistematik, acak, distance geometry, Monte Carlo, dan ensemble tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

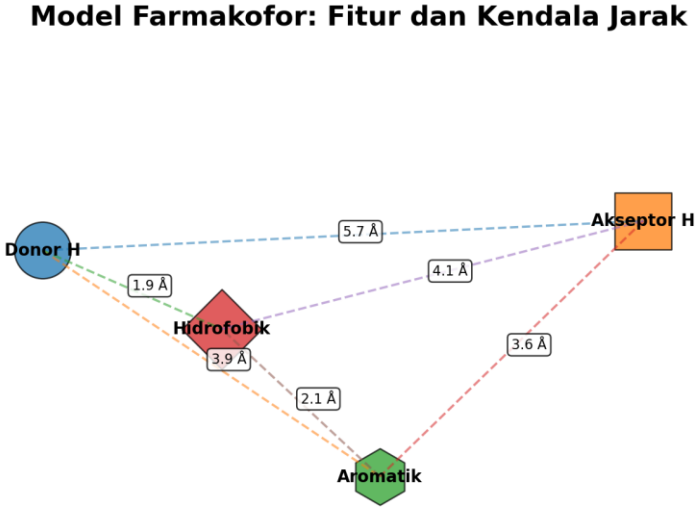
Kesalahan yang sering muncul adalah menghasilkan terlalu sedikit konformer atau tidak melakukan deduplikasi RMSD. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan jumlah dan keragaman konformer dari dua metode. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap sistematik, acak, distance geometry, Monte Carlo, dan ensemble harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 3.2. Model farmakofor sebagai susunan fitur dan kendala jarak.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 3.4 algoritme pencarian konformer perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti sistematik dan hubungan operasionalnya dengan acak. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu metode sistematik, acak, Monte Carlo, dinamika molekul, distance geometry, dan algoritme evolusioner memiliki strategi eksplorasi berbeda, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan sistematik mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar distance geometry. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah acak, distance geometry, dan Monte Carlo harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan

sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan sistematik–ensemble mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan sistematik, acak, distance geometry, Monte Carlo, ensemble sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 3.4 algoritme pencarian konformer perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap

transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi sistematik, atau konversi acak—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan distance geometry sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menghasilkan terlalu sedikit konformer atau tidak melakukan deduplikasi RMSD; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 3.4 algoritme pencarian konformer dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi

yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.4 algoritme pencarian konformer, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai sistematik–ensemble, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas membandingkan jumlah dan keragaman konformer dari dua metode sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok

perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.4 algoritme pencarian konformer sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang sistematik akan memengaruhi cara acak dihitung, cara distance geometry dibandingkan, dan cara Monte Carlo dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap

memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi sistematik, acak, distance geometry, Monte Carlo, ensemble dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
sistematik	Bagaimana sistematik didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.5 Penyelarasan, RMSD, dan Kemiripan Bentuk

Perbandingan 3d memerlukan definisi korespondensi atom, penyelarasan optimal, serta metrik seperti rmsd, overlap volume, dan kemiripan medan. Pokok bahasan ini menempatkan alignment, RMSD, shape overlap, korespondensi atom, dan medan sebagai komponen yang harus dipahami secara terpadu. Dalam informatika

kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah alignment, RMSD, shape overlap, korespondensi atom, dan medan tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah penyelarasan dua analog yang memiliki scaffold sama. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menginterpretasikan RMSD kecil sebagai kesamaan aktivitas tanpa memperhatikan fitur kimia. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada

menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menghitung RMSD setelah dan sebelum alignment. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap alignment, RMSD, shape overlap, korespondensi atom, dan medan harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
alignment	Peran alignment pada konteks Penyelarasan, RMSD, dan Kemiripan Bentuk.
RMSD	Peran RMSD pada konteks Penyelarasan, RMSD, dan Kemiripan Bentuk.
shape overlap	Peran shape overlap pada konteks Penyelarasan, RMSD, dan Kemiripan Bentuk.
korespondensi atom	Peran korespondensi atom pada konteks Penyelarasan, RMSD, dan Kemiripan Bentuk.
medan	Peran medan pada konteks Penyelarasan, RMSD, dan Kemiripan Bentuk.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus penyelarasan dua analog yang memiliki scaffold sama?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Selanjutnya, unit analisis pada 3.5 penyelarasan, rmsd, dan kemiripan bentuk perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti alignment dan hubungan operasionalnya dengan RMSD. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada sisi lain, representasi selalu bersifat selektif. Fokus bagian ini, yaitu perbandingan 3D memerlukan definisi korespondensi atom, penyelarasan optimal, serta metrik seperti RMSD, overlap volume, dan kemiripan medan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan alignment mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar shape overlap. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam analisis yang lebih rinci, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah RMSD, shape overlap, dan korespondensi atom harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus penyelarasan dua analog yang memiliki scaffold sama, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif kimia, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan alignment–medan mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif komputasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan alignment, RMSD, shape overlap, korespondensi atom, medan

sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada tahap verifikasi, sumber data untuk 3.5 penyelarasan, rmsd, dan kemiripan bentuk perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk penyelarasan dua analog yang memiliki scaffold sama, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Untuk pembelajaran berbasis proyek, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi alignment, atau konversi RMSD—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa

bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam skenario industri, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan shape overlap sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam riset akademik, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menginterpretasikan RMSD kecil sebagai kesamaan aktivitas tanpa memperhatikan fitur kimia; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pertama, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Selanjutnya, algoritme pada 3.5 penyelarasan, rmsd, dan kemiripan bentuk dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk penyelarasan dua analog yang memiliki scaffold sama, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada sisi lain, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam analisis yang lebih rinci, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif kimia, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif komputasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada tahap verifikasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.5 penyelarasan, rmsd, dan kemiripan bentuk, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Untuk pembelajaran berbasis proyek, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk penyelarasan dua analog yang memiliki scaffold sama, mahasiswa perlu menjelaskan unit independensi:

apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam skenario industri, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam riset akademik, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai alignment–medan, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pertama, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Selanjutnya, studi kasus penyelarasan dua analog yang memiliki scaffold sama dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada sisi lain, aktivitas menghitung RMSD setelah dan sebelum alignment sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam analisis yang lebih rinci, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif kimia, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif komputasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.5 penyelarasan, rmsd, dan kemiripan bentuk sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang alignment akan memengaruhi cara RMSD dihitung, cara shape overlap dibandingkan, dan cara korespondensi atom dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada penyelarasan dua analog yang memiliki scaffold sama', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi alignment, RMSD, shape overlap, korespondensi atom, medan dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus

entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
alignment	Bagaimana alignment didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
medan	Bagaimana medan memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.6 Farmakofor dan Kueri 3D

Farmakofor merangkum susunan spasial fitur donor, akseptor, aromatik, bermuatan, dan hidrofobik yang dianggap penting bagi interaksi. Pokok bahasan ini menempatkan donor, akseptor, aromatik, hidrofobik, dan kendala jarak sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah donor, akseptor, aromatik, hidrofobik, dan kendala jarak tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan

atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah model tiga titik untuk kandidat ligan yang berbagi aktivitas. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyamakan farmakofor dengan satu struktur kimia tertentu. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak

menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah merancang hipotesis farmakofor dari beberapa ligan aktif. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap donor, akseptor, aromatik, hidrofobik, dan kendala jarak harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus model tiga titik untuk kandidat ligan yang berbagi aktivitas?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 3.6 farmakofor dan kueri 3d perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti donor dan hubungan operasionalnya dengan akseptor. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu farmakofor merangkum susunan spasial fitur donor, akseptor, aromatik, bermuatan, dan hidrofobik yang dianggap penting bagi interaksi, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan donor mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar aromatik. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana

yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah akseptor, aromatik, dan hidrofobik harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus model tiga titik untuk kandidat ligan yang berbagi aktivitas, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan donor–kendala jarak mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan donor, akseptor, aromatik, hidrofobik, kendala jarak sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 3.6 farmakofor dan kueri 3d perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk model tiga titik untuk kandidat ligan yang berbagi aktivitas, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi donor, atau konversi akseptor—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan aromatik sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyamakan farmakofor dengan

satu struktur kimia tertentu; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 3.6 farmakofor dan kueri 3d dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk model tiga titik untuk kandidat ligan yang berbagi aktivitas, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas

dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.6 farmakofor dan kueri 3d, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk model tiga titik untuk kandidat ligan yang berbagi aktivitas, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai donor–kendala jarak, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman

eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus model tiga titik untuk kandidat ligan yang berbagi aktivitas dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas merancang hipotesis farmakofor dari beberapa ligan aktif sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.6 farmakofor dan kueri 3d sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang donor akan memengaruhi cara akseptor dihitung, cara aromatik dibandingkan, dan cara hidrofobik dimasukkan ke

model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada model tiga titik untuk kandidat ligan yang berbagi aktivitas', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi donor, akseptor, aromatik, hidrofobik, kendala jarak dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
donor	Bagaimana donor didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
kendala jarak	Bagaimana kendala jarak memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.7 Basis Data 3D Fleksibel dan Pencarian

Pencarian 3d dapat menyimpan banyak konformer atau membangkitkan konformer saat kueri, dengan trade-off antara penyimpanan, waktu, dan sensitivitas. Pokok bahasan ini menempatkan basis data konformer, indeks 3D, kunci jarak, fleksibilitas, dan sensitivitas sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah basis data konformer, indeks 3D, kunci jarak, fleksibilitas, dan sensitivitas tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengabaikan protonasi dan tautomer yang mengubah fitur farmakofor. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menetapkan toleransi jarak dan menganalisis pengaruhnya. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap basis data konformer, indeks 3D, kunci jarak, fleksibilitas, dan sensitivitas harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
basis data konformer	Peran basis data konformer pada konteks Basis Data 3D Fleksibel dan Pencarian.
indeks 3D	Peran indeks 3D pada konteks Basis Data 3D Fleksibel dan Pencarian.

Konsep	Makna operasional dalam bagian ini
kunci jarak	Peran kunci jarak pada konteks Basis Data 3D Fleksibel dan Pencarian.
fleksibilitas	Peran fleksibilitas pada konteks Basis Data 3D Fleksibel dan Pencarian.
sensitivitas	Peran sensitivitas pada konteks Basis Data 3D Fleksibel dan Pencarian.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 3.7 basis data 3d fleksibel dan pencarian perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti basis data konformer dan hubungan operasionalnya dengan indeks 3D. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu pencarian 3D dapat menyimpan banyak konformer atau membangkitkan konformer saat kueri, dengan trade-off antara penyimpanan, waktu, dan sensitivitas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan basis data konformer mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar kunci jarak. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah indeks 3D, kunci jarak, dan fleksibilitas harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu, satu

nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan basis data konformer–sensitivitas mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan basis data konformer, indeks 3D, kunci jarak, fleksibilitas, sensitivitas sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 3.7 basis data 3d fleksibel dan pencarian perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau

pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi basis data konformer, atau konversi indeks 3D—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan kunci jarak sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengabaikan protonasi dan tautomer yang mengubah fitur farmakofor; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan

format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 3.7 basis data 3d fleksibel dan pencarian dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi

pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.7 basis data 3d fleksibel dan pencarian, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai basis data konformer–sensitivitas, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas menetapkan toleransi jarak dan menganalisis pengaruhnya sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.7 basis data 3d fleksibel dan pencarian sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang basis data konformer akan memengaruhi cara indeks 3D dihitung, cara kunci jarak dibandingkan, dan cara fleksibilitas dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi basis data konformer, indeks 3D, kunci jarak, fleksibilitas, sensitivitas dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
basis data konformer	Bagaimana basis data konformer didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sensitivitas	Bagaimana sensitivitas memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

3.8 Visualisasi Molekuler sebagai Alat Analisis

Visualisasi bukan sekadar dekorasi; ia harus mendukung inspeksi geometri, kualitas data, interaksi, ketidakpastian, dan komunikasi keputusan. Pokok bahasan ini menempatkan rendering, representasi stik, permukaan, warna atom, dan annotasi sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah rendering, representasi stik, permukaan, warna atom, dan annotasi tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan perspektif dan warna yang menyesatkan tanpa legenda. Kekeliruan tersebut dapat menghasilkan analisis yang

tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat gambar molekul dengan caption dan sumber yang lengkap. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap rendering, representasi stik, permukaan, warna atom, dan annotasi harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 3.8 visualisasi molekuler sebagai alat analisis perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti rendering dan hubungan operasionalnya dengan representasi stik. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu visualisasi bukan sekadar dekorasi; ia harus mendukung inspeksi geometri, kualitas data, interaksi, ketidakpastian, dan komunikasi keputusan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan rendering mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar permukaan. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah representasi stik, permukaan, dan warna atom harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan rendering–annotasi mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan rendering, representasi stik, permukaan, warna atom, annotasi sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 3.8 visualisasi molekuler sebagai alat analisis perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi rendering, atau konversi representasi stik—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat

berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan permukaan sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan perspektif dan warna yang menyesatkan tanpa legenda; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 3.8 visualisasi molekuler sebagai alat analisis dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan

eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 3.8 visualisasi molekuler sebagai alat analisis, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein, mahasiswa perlu menjelaskan unit independensi: apakah

tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai rendering–annotasi, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membuat gambar molekul dengan caption dan sumber yang lengkap sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 3.8 visualisasi molekuler sebagai alat analisis sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang rendering akan memengaruhi cara representasi stik dihitung, cara permukaan dibandingkan, dan cara warna atom dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi rendering, representasi stik, permukaan, warna atom, annotasi dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus

entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
rendering	Bagaimana rendering didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
annotasi	Bagaimana annotasi memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$\mathrm{RMSD} = \sqrt{[(1/N)\,\Sigma_i \,\|r_i - r_i'\|^2]}$$

RMSD

Praktikum 3 — Generasi dan Analisis Konformer

Membangkitkan ensemble konformer ETKDG, optimasi gaya gaya, deduplikasi RMSD, dan visualisasi distribusi energi.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.

- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from rdkit import Chem
from rdkit.Chem import AllChem

mol = Chem.AddHs(Chem.MolFromSmiles("CCCC"))
ids = AllChem.EmbedMultipleConfs(mol, numConfs=50,
                                 params=AllChem.ETKDGv3())
energies = []
for cid in ids:
    AllChem.MMFFOptimizeMolecule(mol, confId=cid)
    props = AllChem.MMFFGetMoleculeProperties(mol)
    ff = AllChem.MMFFGetMoleculeForceField(mol, props, confId=cid)
    energies.append((cid, ff.CalcEnergy()))
print(sorted(energies, key=lambda x: x[1])[:5])
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?

- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Struktur Molekul 3D harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara koordinat dan ketidakpastian pada konteks konformasi ligan dalam kristal protein dibandingkan konformasi larutan.
2. Identifikasi minimal tiga risiko jika peneliti menganggap satu struktur kristal sebagai satu-satunya konformasi biologis.
3. Jelaskan dengan contoh hubungan antara Kartesian dan transformasi pada konteks rotasi ikatan tunggal mengubah dihedral tanpa mengubah konektivitas.
4. Identifikasi minimal tiga risiko jika peneliti membandingkan struktur tanpa penyelarasan atau invariansi rotasi-translasi.
5. Jelaskan dengan contoh hubungan antara minimum energi dan fleksibilitas pada konteks butana memiliki konformasi anti dan gauche dengan energi berbeda.
6. Identifikasi minimal tiga risiko jika peneliti memilih konformer energi terendah di vakum sebagai konformer aktif tanpa bukti.
7. Jelaskan dengan contoh hubungan antara sistematik dan ensemble pada konteks RDKit ETKDG menghasilkan ensemble yang kemudian dioptimasi.
8. Identifikasi minimal tiga risiko jika peneliti menghasilkan terlalu sedikit konformer atau tidak melakukan deduplikasi RMSD.
9. Jelaskan dengan contoh hubungan antara alignment dan medan pada konteks penyelarasan dua analog yang memiliki scaffold sama.

10. Identifikasi minimal tiga risiko jika peneliti menginterpretasikan RMSD kecil sebagai kesamaan aktivitas tanpa memperhatikan fitur kimia.
11. Jelaskan dengan contoh hubungan antara donor dan kendala jarak pada konteks model tiga titik untuk kandidat ligan yang berbagi aktivitas.
12. Identifikasi minimal tiga risiko jika peneliti menyamakan farmakofor dengan satu struktur kimia tertentu.
13. Jelaskan dengan contoh hubungan antara basis data konformer dan sensitivitas pada konteks pencarian molekul yang memenuhi tiga jarak fitur dalam toleransi tertentu.
14. Identifikasi minimal tiga risiko jika peneliti mengabaikan protonasi dan tautomer yang mengubah fitur farmakofor.
15. Jelaskan dengan contoh hubungan antara rendering dan annotasi pada konteks menampilkan ikatan hidrogen dan benturan sterik pada kompleks ligan-protein.
16. Identifikasi minimal tiga risiko jika peneliti menggunakan perspektif dan warna yang menyesatkan tanpa legenda.
17. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
18. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
19. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
20. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 4. BASIS DATA KIMIA, PENCARIAN, KURASI, DAN INTEROPERABILITAS

Orientasi Bab

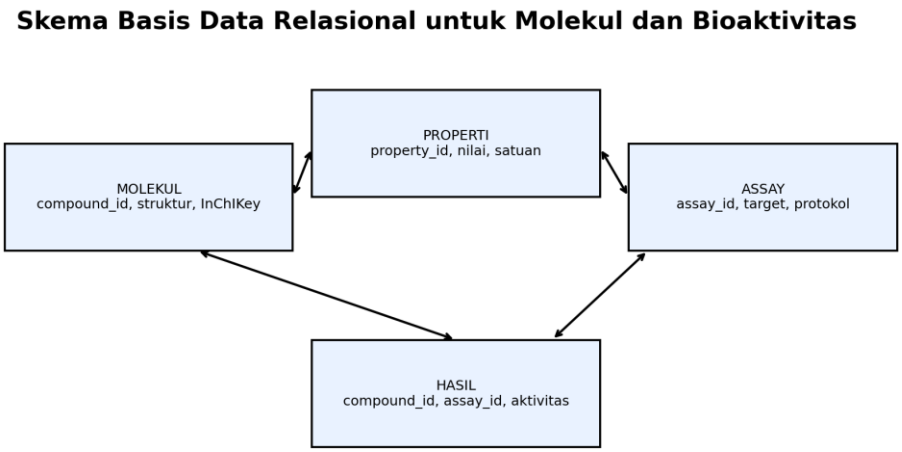
Bab ini menyusun fondasi konseptual dan praktis mengenai basis data dan kurasi. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti basis data dan kurasi dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

accuracy, API, candidate set, CAS, ChEBI, ChEMBL, columnar, completeness, consistency, controlled vocabulary, deduplikasi, document store, fingerprint, fitness for purpose, foreign key, graph database, identity, indeks, join, JSON, latency, normalisasi, ontology, pagination, parent, primary key, PubChem, query, rate limit, RDF, REST, salt stripping, selectivity, semantic web, skema, standardization, tabel, triple, uncertainty, ZINC.



Gambar 4.1. Pemisahan entitas molekul, properti, assay, dan hasil pada model relasional.

4.1 Model Data Relasional dan Normalisasi

Basis data relasional memisahkan entitas molekul, sampel, assay, hasil, target, dan sumber untuk mengurangi redundansi dan menjaga integritas. Pokok bahasan ini menempatkan tabel, primary key, foreign key, normalisasi, dan join sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah tabel, primary key, foreign key, normalisasi, dan join tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah hubungan compound_id, assay_id, dan activity_value dalam tabel hasil. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyimpan seluruh informasi dalam satu tabel lebar yang sulit dirawat. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah merancang skema relasional sederhana untuk data praktikum. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap tabel, primary key, foreign key, normalisasi, dan join harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia

adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
tabel	Peran tabel pada konteks Model Data Relasional dan Normalisasi.
primary key	Peran primary key pada konteks Model Data Relasional dan Normalisasi.
foreign key	Peran foreign key pada konteks Model Data Relasional dan Normalisasi.
normalisasi	Peran normalisasi pada konteks Model Data Relasional dan Normalisasi.
join	Peran join pada konteks Model Data Relasional dan Normalisasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus hubungan compound_id, assay_id, dan activity_value dalam tabel hasil?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 4.1 model data relasional dan normalisasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti tabel dan hubungan operasionalnya dengan primary key. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu basis data relasional memisahkan entitas molekul, sampel, assay, hasil, target, dan sumber untuk mengurangi redundansi dan menjaga integritas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan tabel mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar foreign key. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah primary key, foreign key, dan normalisasi harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus hubungan compound_id, assay_id, dan activity_value dalam tabel hasil, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan tabel–join mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan tabel, primary key, foreign key, normalisasi, join sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 4.1 model data relasional dan normalisasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk hubungan compound_id, assay_id, dan activity_value dalam tabel hasil, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi tabel, atau konversi primary key—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan foreign key sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyimpan seluruh informasi dalam satu tabel lebar yang sulit dirawat; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier

yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 4.1 model data relasional dan normalisasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk hubungan compound_id, assay_id, dan activity_value dalam tabel hasil, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.1 model data relasional dan normalisasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk hubungan compound_id, assay_id, dan activity_value dalam tabel hasil, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai tabel–join, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus hubungan compound_id, assay_id, dan activity_value dalam tabel hasil dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas merancang skema relasional sederhana untuk data praktikum sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.1 model data relasional dan normalisasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang tabel akan memengaruhi cara primary key dihitung, cara foreign key dibandingkan, dan cara normalisasi dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada hubungan compound_id, assay_id, dan activity_value dalam tabel hasil', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi tabel, primary key, foreign key, normalisasi, join dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
tabel	Bagaimana tabel didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
join	Bagaimana join memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.2 Basis Data Graf, Dokumen, dan Kolom

Model nonrelasional berguna ketika hubungan kompleks, skema berubah cepat, atau volume besar, tetapi pemilihannya harus mengikuti pola kueri dan konsistensi yang dibutuhkan. Pokok bahasan ini menempatkan graph database, document store, columnar, skema, dan query sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah graph database, document store, columnar, skema, dan query tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah jejaring senyawa-target-penyakit sebagai knowledge graph. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih teknologi karena populer tanpa mengevaluasi kebutuhan transaksi. Kekeliruan tersebut dapat menghasilkan

analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan representasi satu dataset pada dua model basis data. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap graph database, document store, columnar, skema, dan query harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus jejaring senyawa-target-penyakit sebagai knowledge graph?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 4.2 basis data graf, dokumen, dan kolom perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti graph database dan hubungan operasionalnya dengan document store. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model nonrelasional berguna ketika hubungan kompleks, skema berubah cepat, atau volume besar, tetapi pemilihannya harus mengikuti pola kueri dan konsistensi yang dibutuhkan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan graph database mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar columnar. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah document store, columnar, dan skema harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus jejaring senyawa-target-penyakit sebagai knowledge graph, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan graph database–query mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan graph database, document store, columnar, skema, query sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 4.2 basis data graf, dokumen, dan kolom perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk jejaring senyawa-target-penyakit sebagai knowledge graph, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi graph database, atau konversi document store—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel

audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan columnar sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih teknologi karena populer tanpa mengevaluasi kebutuhan transaksi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 4.2 basis data graf, dokumen, dan kolom dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk jejaring senyawa-target-penyakit sebagai knowledge graph, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.2 basis data graf, dokumen, dan kolom, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk jejaring senyawa-target-penyakit sebagai knowledge graph, mahasiswa perlu menjelaskan unit independensi: apakah

tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai graph database–query, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus jejaring senyawa-target-penyakit sebagai knowledge graph dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas membandingkan representasi satu dataset pada dua model basis data sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.2 basis data graf, dokumen, dan kolom sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang graph database akan memengaruhi cara document store dihitung, cara columnar dibandingkan, dan cara skema dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada jejaring senyawa-target-penyakit sebagai knowledge graph', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi graph database, document store, columnar, skema, query dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas,

membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
graph database	Bagaimana graph database didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
query	Bagaimana query memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.3 Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri

Indeks kimia mempercepat pencarian struktur dengan menyimpan fitur yang dapat menyingkirkan kandidat sebelum pencocokan rinci. Pokok bahasan ini menempatkan indeks, fingerprint, selectivity, candidate set, dan latency sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah indeks, fingerprint, selectivity, candidate set, dan latency tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis.

Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah bit tertentu wajib muncul untuk kueri substruktur. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengoptimasi kecepatan dengan mengorbankan kebenaran hasil. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengukur waktu kueri sebelum dan sesudah indeks. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap indeks, fingerprint, selectivity, candidate set, dan latency harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
indeks	Peran indeks pada konteks Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri.
fingerprint	Peran fingerprint pada konteks Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri.
selectivity	Peran selectivity pada konteks Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri.
candidate set	Peran candidate set pada konteks Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri.
latency	Peran latency pada konteks Indeks Struktur, Fingerprint Screening, dan Kinerja Kueri.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus bit tertentu wajib muncul untuk kueri substruktur?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 4.3 indeks struktur, fingerprint screening, dan kinerja kueri perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti indeks dan hubungan operasionalnya dengan fingerprint. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu indeks kimia mempercepat pencarian struktur dengan menyimpan fitur yang dapat menyingkirkan kandidat sebelum pencocokan rinci, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan indeks mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar selectivity. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah fingerprint, selectivity, dan candidate set harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus bit tertentu wajib muncul untuk kueri substruktur, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan indeks–latency mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan indeks, fingerprint, selectivity, candidate set, latency sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 4.3 indeks struktur, fingerprint screening, dan kinerja kueri perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk bit tertentu wajib muncul untuk kueri substruktur, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi indeks, atau konversi fingerprint—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan selectivity sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci

sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengoptimasi kecepatan dengan mengorbankan kebenaran hasil; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 4.3 indeks struktur, fingerprint screening, dan kinerja kueri dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan

keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk bit tertentu wajib muncul untuk kueri substruktur, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.3 indeks struktur, fingerprint screening, dan kinerja kueri, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk bit tertentu wajib muncul untuk kueri substruktur, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai indeks–latency, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus bit tertentu wajib muncul untuk kueri substruktur dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas mengukur waktu kueri sebelum dan sesudah indeks sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.3 indeks struktur, fingerprint screening, dan kinerja kueri sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang indeks akan memengaruhi cara fingerprint dihitung, cara selectivity dibandingkan, dan cara candidate set dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada bit tertentu wajib muncul untuk kueri substruktur', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi indeks, fingerprint, selectivity, candidate set, latency dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
indeks	Bagaimana indeks didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
latency	Bagaimana latency memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.4 Standardisasi, Deduplikasi, dan Rekonsiliasi Identitas

Rekonsiliasi data harus membedakan parent compound, garam, solvat, isotop, stereoisomer, tautomer, dan campuran sesuai pertanyaan ilmiah. Pokok bahasan ini menempatkan parent, salt stripping, deduplikasi, identity, dan standardization sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah parent, salt stripping, deduplikasi, identity, dan standardization tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggabungkan rekaman hanya berdasarkan nama teks. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

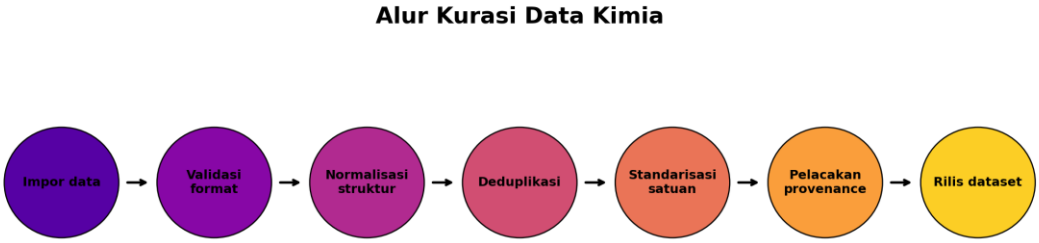
Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat aturan deduplikasi bertingkat menggunakan InChIKey dan metadata. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap parent, salt stripping, deduplikasi, identity, dan standardization harus selalu dihubungkan dengan bab lain.

Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 4.2. Tahapan kurasi dan pelepasan dataset yang dapat ditelusuri.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 4.4 standardisasi, deduplikasi, dan rekonsiliasi identitas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti parent dan hubungan operasionalnya dengan salt stripping. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu rekonsiliasi data harus membedakan parent compound, garam, solvat, isotop, stereoisomer, tautomer, dan campuran sesuai pertanyaan ilmiah, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan parent mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar deduplikasi. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi,

dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah salt stripping, deduplikasi, dan identity harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan parent–standardization mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan parent, salt stripping, deduplikasi, identity, standardization sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 4.4 standardisasi, deduplikasi, dan rekonsiliasi identitas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi parent, atau konversi salt stripping—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan deduplikasi sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier,

struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggabungkan rekaman hanya berdasarkan nama teks; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 4.4 standardisasi, deduplikasi, dan rekonsiliasi identitas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.4 standardisasi, deduplikasi, dan rekonsiliasi identitas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual,

confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai parent–standardization, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas membuat aturan deduplikasi bertingkat menggunakan InChIKey dan metadata sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.4 standardisasi, deduplikasi, dan rekonsiliasi identitas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang parent akan memengaruhi cara salt stripping dihitung, cara deduplikasi dibandingkan, dan cara identity dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi parent, salt stripping, deduplikasi, identity, standardization dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
parent	Bagaimana parent didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
standardization	Bagaimana standardization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.5 Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian

Kualitas data bersifat multidimensi dan harus dinilai terhadap tujuan penggunaan, bukan hanya jumlah rekaman. Pokok bahasan ini menempatkan completeness, accuracy, consistency, uncertainty, dan fitness for purpose sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah completeness, accuracy, consistency, uncertainty, dan fitness for purpose tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting

dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menghapus outlier tanpa menyelidiki penyebab kimia atau eksperimen. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat profil kualitas untuk 100 rekaman. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap completeness, accuracy, consistency, uncertainty, dan fitness for purpose harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
completeness	Peran completeness pada konteks Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian.
accuracy	Peran accuracy pada konteks Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian.
consistency	Peran consistency pada konteks Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian.
uncertainty	Peran uncertainty pada konteks Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian.
fitness for purpose	Peran fitness for purpose pada konteks Kualitas Data: Kelengkapan, Ketepatan, Konsistensi, dan Ketidakpastian.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 4.5 kualitas data: kelengkapan, ketepatan, konsistensi, dan ketidakpastian perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti completeness dan hubungan operasionalnya dengan accuracy. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu kualitas data bersifat multidimensi dan harus dinilai terhadap tujuan penggunaan,

bukan hanya jumlah rekaman, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan completeness mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar consistency. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah accuracy, consistency, dan uncertainty harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan completeness–fitness for purpose mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan completeness, accuracy, consistency, uncertainty, fitness for purpose sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 4.5 kualitas data: kelengkapan, ketepatan, konsistensi, dan ketidakpastian perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi completeness, atau konversi accuracy—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan consistency sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menghapus outlier tanpa menyelidiki penyebab kimia atau eksperimen; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 4.5 kualitas data: kelengkapan, ketepatan, konsistensi, dan ketidakpastian dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan, mahasiswa perlu menuliskan pseudocode ringkas sebelum

menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.5 kualitas data: kelengkapan, ketepatan, konsistensi, dan ketidakpastian, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai completeness–fitness for purpose, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas membuat profil kualitas untuk 100 rekaman sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.5 kualitas data: kelengkapan, ketepatan, konsistensi, dan ketidakpastian sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang completeness akan memengaruhi cara accuracy dihitung, cara consistency dibandingkan, dan cara uncertainty dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi completeness, accuracy, consistency, uncertainty, fitness for purpose dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas

kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
completeness	Bagaimana completeness didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
fitness for purpose	Bagaimana fitness for purpose memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.6 Ontologi, Taksonomi, dan Data Tertaut

Ontologi menyediakan makna formal untuk entitas dan relasi sehingga data dari sumber berbeda dapat diintegrasikan dengan interpretasi yang lebih konsisten. Pokok bahasan ini menempatkan ontology, RDF, triple, controlled vocabulary, dan semantic web sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah ontology, RDF, triple, controlled vocabulary, dan semantic web tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika

hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan istilah bebas yang berbeda untuk konsep sama. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengubah tiga pernyataan kimia menjadi triple subjek-predikat-objek. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap ontology, RDF, triple, controlled vocabulary, dan semantic web harus selalu dihubungkan dengan bab lain.

Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 4.6 ontologi, taksonomi, dan data tertaut perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti ontology dan hubungan operasionalnya dengan RDF. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu ontologi menyediakan makna formal untuk entitas dan relasi sehingga data dari sumber berbeda dapat diintegrasikan dengan interpretasi yang lebih konsisten, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan ontology mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar triple. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah RDF, triple, dan controlled vocabulary harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata

bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan ontology–semantic web mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan ontology, RDF, triple, controlled vocabulary, semantic web sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 4.6 ontologi, taksonomi, dan data tertaut perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder,

memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi ontology, atau konversi RDF—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan triple sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan istilah bebas yang berbeda untuk konsep sama; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 4.6 ontologi, taksonomi, dan data tertaut dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap

akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.6

ontologi, taksonomi, dan data tertaut, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai ontology–semantic web, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas mengubah tiga pernyataan kimia menjadi triple subjek-predikat-objek sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.6 ontologi, taksonomi, dan data tertaut sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang ontology akan memengaruhi cara RDF dihitung, cara triple dibandingkan, dan cara controlled vocabulary dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu', melainkan membandingkan dua kebijakan representasi atau dua strategi

validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi ontology, RDF, triple, controlled vocabulary, semantic web dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ontology	Bagaimana ontology didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
semantic web	Bagaimana semantic web memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.7 Sumber Data Publik dan Komersial

Sumber data memiliki cakupan, lisensi, kualitas kurasi, mekanisme akses, dan frekuensi pembaruan yang berbeda; pemilihan harus disertai evaluasi sumber.

Pokok bahasan ini menempatkan PubChem, ChEMBL, ChEBI, ZINC, dan CAS sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah PubChem, ChEMBL, ChEBI, ZINC, dan CAS tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menggabungkan struktur PubChem dengan bioaktivitas ChEMBL. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap semua rekaman publik telah diverifikasi secara identik. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh,

menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan provenance satu senyawa pada dua basis data. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap PubChem, ChEMBL, ChEBI, ZINC, dan CAS harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
PubChem	Peran PubChem pada konteks Sumber Data Publik dan Komersial.
ChEMBL	Peran ChEMBL pada konteks Sumber Data Publik dan Komersial.
ChEBI	Peran ChEBI pada konteks Sumber Data Publik dan Komersial.
ZINC	Peran ZINC pada konteks Sumber Data Publik dan Komersial.
CAS	Peran CAS pada konteks Sumber Data Publik dan Komersial.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menggabungkan struktur PubChem dengan bioaktivitas ChEMBL?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 4.7 sumber data publik dan komersial perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti PubChem dan hubungan operasionalnya dengan ChEMBL. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu sumber data memiliki cakupan, lisensi, kualitas kurasi, mekanisme akses, dan frekuensi pembaruan yang berbeda; pemilihan harus disertai evaluasi sumber, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan PubChem mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar ChEBI. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah ChEMBL, ChEBI, dan ZINC harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menggabungkan struktur PubChem dengan bioaktivitas ChEMBL, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan PubChem–CAS mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan PubChem, ChEMBL, ChEBI, ZINC, CAS sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 4.7 sumber data publik dan komersial perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menggabungkan struktur PubChem dengan bioaktivitas ChEMBL, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi PubChem, atau konversi ChEMBL—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok

tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan ChEBI sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap semua rekaman publik telah diverifikasi secara identik; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 4.7 sumber data publik dan komersial dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menggabungkan struktur PubChem dengan bioaktivitas ChEMBL, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.7 sumber data publik dan komersial, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menggabungkan struktur PubChem dengan bioaktivitas ChEMBL, mahasiswa perlu menjelaskan unit independensi: apakah

tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai PubChem–CAS, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus menggabungkan struktur PubChem dengan bioaktivitas ChEMBL dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas membandingkan provenance satu senyawa pada dua basis data sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.7 sumber data publik dan komersial sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang PubChem akan memengaruhi cara ChEMBL dihitung, cara ChEBI dibandingkan, dan cara ZINC dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menggabungkan struktur PubChem dengan bioaktivitas ChEMBL', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi PubChem, ChEMBL, ChEBI, ZINC, CAS dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun

skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PubChem	Bagaimana PubChem didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
CAS	Bagaimana CAS memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

4.8 API, Layanan Web, dan Otomasi Akuisisi Data

Api memungkinkan kueri terprogram, pagination, filter, dan integrasi pipeline, tetapi memerlukan pengendalian versi, cache, rate limit, dan log permintaan. Pokok bahasan ini menempatkan REST, API, JSON, pagination, dan rate limit sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah REST, API, JSON, pagination, dan rate limit tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis.

Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mengambil properti beberapa CID dari PUG-REST. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengunduh data tanpa merekam tanggal, parameter, dan versi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis pseudocode pengambilan data yang dapat diulang. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap REST, API, JSON, pagination, dan rate limit harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mengambil properti beberapa CID dari PUG-REST?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 4.8 api, layanan web, dan otomasi akuisisi data perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti REST dan hubungan operasionalnya dengan API. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu API memungkinkan kueri terprogram, pagination, filter, dan integrasi pipeline, tetapi memerlukan pengendalian versi, cache, rate limit, dan log permintaan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan REST mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar JSON. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah API, JSON, dan pagination harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mengambil properti beberapa CID dari PUG-REST, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan REST–rate limit mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan REST, API, JSON, pagination, rate limit sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 4.8 api, layanan web, dan otomasi akuisisi data perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mengambil properti beberapa CID dari PUG-REST, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi REST, atau konversi API—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan JSON sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengunduh data tanpa merekam tanggal, parameter, dan versi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier

yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 4.8 api, layanan web, dan otomasi akuisisi data dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mengambil properti beberapa CID dari PUG-REST, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 4.8 api, layanan web, dan otomasi akuisisi data, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mengambil properti beberapa CID dari PUG-REST, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai REST–rate limit, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus mengambil properti beberapa CID dari PUG-REST dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas menulis pseudocode pengambilan data yang dapat diulang sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama,

mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 4.8 api, layanan web, dan otomasi akuisisi data sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang REST akan memengaruhi cara API dihitung, cara JSON dibandingkan, dan cara pagination dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mengambil properti beberapa CID dari PUG-REST', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi REST, API, JSON, pagination, rate limit dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
REST	Bagaimana REST didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
rate limit	Bagaimana rate limit memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Praktikum 4 — Basis Data SQLite untuk Senyawa dan Assay

Membuat skema relasional, memasukkan data, menjalankan SQL join, serta mencatat aturan standardisasi dan provenance.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
import sqlite3

con = sqlite3.connect("kimia.db")
cur = con.cursor()
cur.execute("CREATE TABLE IF NOT EXISTS molecule (compound_id TEXT
PRIMARY KEY, smiles TEXT, inchikey TEXT)")
cur.execute("CREATE TABLE IF NOT EXISTS assay (assay_id TEXT PRIMARY KEY,
endpoint TEXT, unit TEXT)")
cur.execute("CREATE TABLE IF NOT EXISTS result (compound_id TEXT,
assay_id TEXT, value REAL)")
con.commit()
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.

- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Basis Data dan Kurasi harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara tabel dan join pada konteks hubungan compound_id, assay_id, dan activity_value dalam tabel hasil.
2. Identifikasi minimal tiga risiko jika peneliti menyimpan seluruh informasi dalam satu tabel lebar yang sulit dirawat.
3. Jelaskan dengan contoh hubungan antara graph database dan query pada konteks jejaring senyawa-target-penyakit sebagai knowledge graph.
4. Identifikasi minimal tiga risiko jika peneliti memilih teknologi karena populer tanpa mengevaluasi kebutuhan transaksi.
5. Jelaskan dengan contoh hubungan antara indeks dan latency pada konteks bit tertentu wajib muncul untuk kueri substruktur.
6. Identifikasi minimal tiga risiko jika peneliti mengoptimasi kecepatan dengan mengorbankan kebenaran hasil.
7. Jelaskan dengan contoh hubungan antara parent dan standardization pada konteks dua pemasok mencantumkan senyawa sama dengan nama dan bentuk garam berbeda.

8. Identifikasi minimal tiga risiko jika peneliti menggabungkan rekaman hanya berdasarkan nama teks.
9. Jelaskan dengan contoh hubungan antara completeness dan fitness for purpose pada konteks nilai IC50 tanpa satuan atau tanpa jenis assay tidak dapat dibandingkan.
10. Identifikasi minimal tiga risiko jika peneliti menghapus outlier tanpa menyelidiki penyebab kimia atau eksperimen.
11. Jelaskan dengan contoh hubungan antara ontology dan semantic web pada konteks relasi senyawa memiliki aktivitas terhadap target dalam assay tertentu.
12. Identifikasi minimal tiga risiko jika peneliti menggunakan istilah bebas yang berbeda untuk konsep sama.
13. Jelaskan dengan contoh hubungan antara PubChem dan CAS pada konteks menggabungkan struktur PubChem dengan bioaktivitas ChEMBL.
14. Identifikasi minimal tiga risiko jika peneliti menganggap semua rekaman publik telah diverifikasi secara identik.
15. Jelaskan dengan contoh hubungan antara REST dan rate limit pada konteks mengambil properti beberapa CID dari PUG-REST.
16. Identifikasi minimal tiga risiko jika peneliti mengunduh data tanpa merekam tanggal, parameter, dan versi.
17. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
18. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
19. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
20. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 5. DESKRIPTOR MOLEKULER, FINGERPRINT, DAN RUANG KIMIA

Orientasi Bab

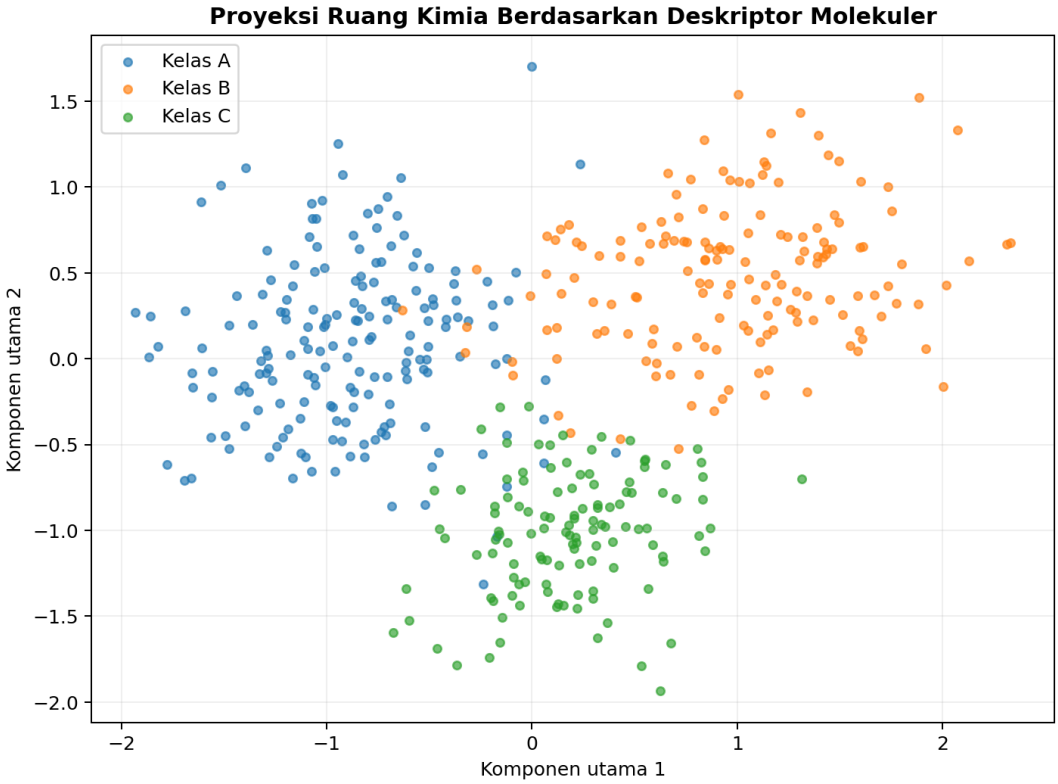
Bab ini menyusun fondasi konseptual dan praktis mengenai deskriptor dan ruang kimia. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti deskriptor dan ruang kimia dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

0D, 1D, 2D, 3D, BCUT, bit collision, connectivity index, correlation filter, CPSA, E-state, ECFP, feature selection, field descriptor, hashing, HBA, HBD, imputation, kappa, leakage, log transform, logP, MACCS, molecular weight, normalization, path fingerprint, PCA, projection, RDF, regularization, shape, SOM, standardization, t-SNE, TPSA, UMAP, variance threshold, vektor fitur, WHIM, Wiener.



Gambar 5.1. Proyeksi dua dimensi dari ruang deskriptor berdimensi lebih tinggi.

5.1 Konsep Deskriptor dan Dimensi Representasi

Deskriptor mengubah molekul menjadi angka atau vektor yang dapat dibandingkan dan dimodelkan; setiap deskriptor membawa asumsi tentang struktur dan sifat yang dianggap relevan. Pokok bahasan ini menempatkan 0D, 1D, 2D, 3D, dan vektor fitur sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam kerangka pengambilan keputusan, istilah 0D, 1D, 2D, 3D, dan vektor fitur tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting

dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan ribuan deskriptor tanpa interpretasi atau kontrol kebocoran. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengelompokkan deskriptor berdasarkan informasi yang dibutuhkan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap 0D, 1D, 2D, 3D, dan vektor fitur harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
0D	Peran 0D pada konteks Konsep Deskriptor dan Dimensi Representasi.
1D	Peran 1D pada konteks Konsep Deskriptor dan Dimensi Representasi.
2D	Peran 2D pada konteks Konsep Deskriptor dan Dimensi Representasi.
3D	Peran 3D pada konteks Konsep Deskriptor dan Dimensi Representasi.
vektor fitur	Peran vektor fitur pada konteks Konsep Deskriptor dan Dimensi Representasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 5.1 konsep deskriptor dan dimensi representasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti 0D dan hubungan operasionalnya dengan 1D. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu deskriptor mengubah molekul menjadi angka atau vektor yang dapat dibandingkan dan dimodelkan; setiap deskriptor membawa asumsi tentang struktur dan sifat yang dianggap relevan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan 0D mungkin sangat efisien untuk pencarian, tetapi belum tentu

memadai untuk menalar 2D. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah 1D, 2D, dan 3D harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan 0D–vektor fitur mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan 0D, 1D, 2D, 3D, vektor fitur sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 5.1 konsep deskriptor dan dimensi representasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi 0D, atau konversi 1D—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan 2D sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur

kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan ribuan deskriptor tanpa interpretasi atau kontrol kebocoran; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 5.1 konsep deskriptor dan dimensi representasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.1 konsep deskriptor dan dimensi representasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai 0D–vektor fitur,

serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas mengelompokkan deskriptor berdasarkan informasi yang dibutuhkan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data

terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.1 konsep deskriptor dan dimensi representasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang 0D akan memengaruhi cara 1D dihitung, cara 2D dibandingkan, dan cara 3D dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi 0D, 1D, 2D, 3D, vektor fitur dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
0D	Bagaimana 0D didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
vektor fitur	Bagaimana vektor fitur memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.2 Deskriptor Konstitusional dan Fisikokimia

Jumlah atom, massa, fraksi heteroatom, logp, polar surface area, dan jumlah ikatan rotasi sering menjadi baseline yang kuat. Pokok bahasan ini menempatkan molecular weight, logP, TPSA, HBD, dan HBA sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah molecular weight, logP, TPSA, HBD, dan HBA tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah profil sederhana untuk menilai kelarutan dan permeabilitas awal. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat

kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap nilai komputasi identik dengan pengukuran eksperimen. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menghitung deskriptor dasar beberapa obat umum. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap molecular weight, logP, TPSA, HBD, dan HBA harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus profil sederhana untuk menilai kelarutan dan permeabilitas awal?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 5.2 deskriptor konstitusional dan fisikokimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti molecular weight dan hubungan operasionalnya dengan logP. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu jumlah atom, massa, fraksi heteroatom, logP, polar surface area, dan jumlah ikatan rotasi sering menjadi baseline yang kuat, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan molecular weight mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar TPSA. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah logP, TPSA, dan HBD harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus profil sederhana untuk menilai kelarutan dan permeabilitas awal, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen

untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan molecular weight–HBA mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan molecular weight, logP, TPSA, HBD, HBA sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 5.2 deskriptor konstitusional dan fisikokimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk profil sederhana untuk menilai kelarutan dan permeabilitas awal, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi molecular weight, atau konversi logP—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan TPSA sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap nilai komputasi identik dengan pengukuran eksperimen; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 5.2 deskriptor konstitusional dan fisikokimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk profil sederhana untuk menilai kelarutan dan permeabilitas awal, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi

keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.2 deskriptor konstitusional dan fisikokimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk profil sederhana untuk menilai kelarutan dan permeabilitas awal, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai molecular weight–HBA, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus profil sederhana untuk menilai kelarutan dan permeabilitas awal dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas menghitung deskriptor dasar beberapa obat umum sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.2 deskriptor konstitusional dan fisikokimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang molecular weight akan memengaruhi cara logP dihitung, cara TPSA dibandingkan, dan cara HBD dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada profil sederhana untuk menilai kelarutan dan permeabilitas awal', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi molecular weight, logP, TPSA, HBD, HBA dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
molecular weight	Bagaimana molecular weight didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
HBA	Bagaimana HBA memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.3 Indeks Topologi dan Deskriptor Graf

Indeks wiener, kier-hall, connectivity, e-state, bcut, dan keluarga lain merangkum konektivitas dengan tingkat interpretabilitas berbeda. Pokok bahasan ini menempatkan Wiener, kappa, connectivity index, E-state, dan BCUT sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang

dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah Wiener, kappa, connectivity index, E-state, dan BCUT tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah membedakan isomer dengan konektivitas berbeda namun rumus molekul sama. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menumpuk deskriptor sangat berkorelasi yang menurunkan stabilitas model. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menghitung korelasi antarindeks topologi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Wiener, kappa, connectivity index, E-state, dan BCUT harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
Wiener	Peran Wiener pada konteks Indeks Topologi dan Deskriptor Graf.
kappa	Peran kappa pada konteks Indeks Topologi dan Deskriptor Graf.
connectivity index	Peran connectivity index pada konteks Indeks Topologi dan Deskriptor Graf.
E-state	Peran E-state pada konteks Indeks Topologi dan Deskriptor Graf.
BCUT	Peran BCUT pada konteks Indeks Topologi dan Deskriptor Graf.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus membedakan isomer dengan konektivitas berbeda namun rumus molekul sama?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 5.3 indeks topologi dan deskriptor graf perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Wiener dan hubungan operasionalnya dengan kappa. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu indeks Wiener, Kier-Hall, connectivity, E-state, BCUT, dan keluarga lain merangkum konektivitas dengan tingkat interpretabilitas berbeda, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Wiener mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar connectivity index. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah kappa, connectivity index, dan E-state harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus membedakan isomer dengan konektivitas berbeda namun rumus molekul sama, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Wiener–BCUT mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan Wiener, kappa, connectivity index, E-state, BCUT sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 5.3 indeks topologi dan deskriptor graf perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk membedakan isomer dengan konektivitas berbeda namun rumus molekul sama, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Wiener, atau konversi kappa—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika

satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan connectivity index sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menumpuk deskriptor sangat berkorelasi yang menurunkan stabilitas model; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 5.3 indeks topologi dan deskriptor graf dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk membedakan isomer dengan konektivitas berbeda namun rumus molekul sama, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang

relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.3 indeks topologi dan deskriptor graf, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk membedakan isomer dengan konektivitas berbeda namun rumus molekul sama, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang

baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Wiener–BCUT, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus membedakan isomer dengan konektivitas berbeda namun rumus molekul sama dapat dirancang sebagai keputusan bertahap.

Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas menghitung korelasi antarindeks topologi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.3 indeks topologi dan deskriptor graf sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Wiener akan memengaruhi cara kappa dihitung, cara connectivity index dibandingkan, dan cara E-state dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada membedakan isomer dengan konektivitas berbeda namun rumus molekul sama', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Wiener, kappa, connectivity index, E-state, BCUT dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Wiener	Bagaimana Wiener didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
BCUT	Bagaimana BCUT memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.4 Fingerprint Struktural dan Circular Fingerprint

Fingerprint berbasis kunci, path, topological torsion, atom pair, dan circular/ecfp mengenkode keberadaan lingkungan atom atau fragmen. Pokok bahasan ini menempatkan MACCS, path fingerprint, ECFP, hashing, dan bit collision sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah MACCS, path fingerprint, ECFP, hashing, dan bit collision tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi.

Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menginterpretasikan bit hashed sebagai fitur tunggal yang pasti. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

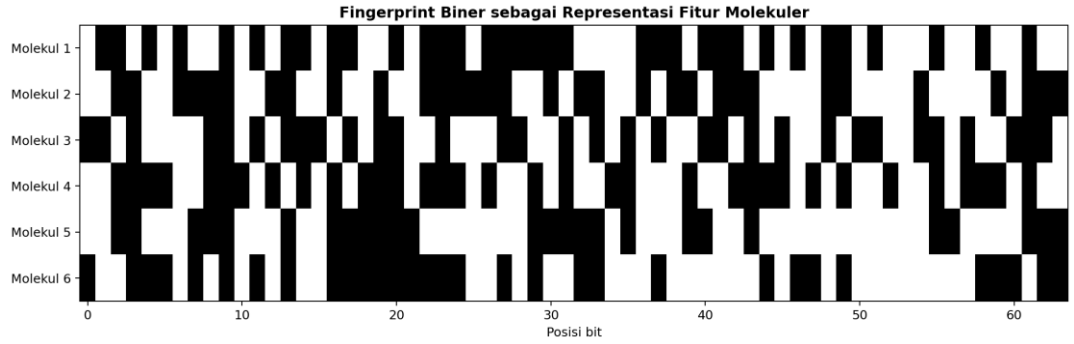
Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan pengaruh radius dan panjang bit. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu

melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap MACCS, path fingerprint, ECFP, hashing, dan bit collision harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 5.2. Fingerprint biner untuk beberapa molekul; setiap baris merupakan vektor fitur.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 5.4 fingerprint struktural dan circular fingerprint perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti MACCS dan hubungan operasionalnya dengan path fingerprint. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu fingerprint berbasis kunci, path, topological torsion, atom pair, dan circular/ECFP

mengenkode keberadaan lingkungan atom atau fragmen, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan MACCS mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar ECFP. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah path fingerprint, ECFP, dan hashing harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan MACCS–bit collision mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan MACCS, path fingerprint, ECFP, hashing, bit collision sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 5.4 fingerprint struktural dan circular fingerprint perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi MACCS, atau konversi path fingerprint—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan ECFP sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menginterpretasikan bit hashed sebagai fitur tunggal yang pasti; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 5.4 fingerprint struktural dan circular fingerprint dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode

memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.4 fingerprint struktural dan circular fingerprint, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai MACCS–bit collision, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas membandingkan pengaruh radius dan panjang bit sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.4 fingerprint struktural dan circular fingerprint sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang MACCS akan memengaruhi cara path fingerprint dihitung, cara ECFP dibandingkan, dan cara hashing dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi MACCS, path fingerprint, ECFP, hashing, bit collision dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MACCS	Bagaimana MACCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
bit collision	Bagaimana bit collision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.5 Deskriptor 3D, Medan, dan Permukaan

Deskriptor 3d memanfaatkan bentuk, distribusi muatan, momen, permukaan, dan medan tetapi peka terhadap konformer serta preparasi. Pokok bahasan ini menempatkan shape, CPSA, WHIM, RDF, dan field descriptor sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah shape, CPSA, WHIM, RDF, dan field descriptor tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah deskriptor permukaan bermuatan untuk prediksi permeabilitas. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menghitung satu konformer tanpa menguji sensitivitas. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan deskriptor dari beberapa konformer. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap shape, CPSA, WHIM, RDF, dan field descriptor harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi

memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
shape	Peran shape pada konteks Deskriptor 3D, Medan, dan Permukaan.
CPSA	Peran CPSA pada konteks Deskriptor 3D, Medan, dan Permukaan.
WHIM	Peran WHIM pada konteks Deskriptor 3D, Medan, dan Permukaan.
RDF	Peran RDF pada konteks Deskriptor 3D, Medan, dan Permukaan.
field descriptor	Peran field descriptor pada konteks Deskriptor 3D, Medan, dan Permukaan.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus deskriptor permukaan bermuatan untuk prediksi permeabilitas?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 5.5 deskriptor 3d, medan, dan permukaan perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti shape dan hubungan operasionalnya dengan CPSA. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu deskriptor 3D memanfaatkan bentuk, distribusi muatan, momen, permukaan, dan medan tetapi peka terhadap konformer serta preparasi, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan shape mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar WHIM. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai

representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah CPSA, WHIM, dan RDF harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus deskriptor permukaan bermuatan untuk prediksi permeabilitas, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan shape–field descriptor mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan shape, CPSA, WHIM, RDF, field descriptor sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 5.5 deskriptor 3d, medan, dan permukaan perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk deskriptor permukaan bermuatan untuk prediksi permeabilitas, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi shape, atau konversi CPSA—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan WHIM sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menghitung satu konformer tanpa

menguji sensitivitas; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 5.5 deskriptor 3d, medan, dan permukaan dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk deskriptor permukaan bermuatan untuk prediksi permeabilitas, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap

penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.5 deskriptor 3d, medan, dan permukaan, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk deskriptor permukaan bermuatan untuk prediksi permeabilitas, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai shape–field descriptor, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan;

ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus deskriptor permukaan bermuatan untuk prediksi permeabilitas dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas membandingkan deskriptor dari beberapa konformer sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.5 deskriptor 3d, medan, dan permukaan sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang shape akan memengaruhi cara CPSA dihitung, cara WHIM dibandingkan, dan cara RDF

dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada deskriptor permukaan bermuatan untuk prediksi permeabilitas', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi shape, CPSA, WHIM, RDF, field descriptor dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
shape	Bagaimana shape didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
field descriptor	Bagaimana field descriptor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.6 Praproses Data: Scaling, Transformasi, dan Missing Value

Fitur dengan rentang berbeda memerlukan scaling; distribusi menceng dan missing value harus ditangani tanpa kebocoran informasi dari data uji. Pokok bahasan ini menempatkan standardization, normalization, log transform, imputation, dan leakage sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah standardization, normalization, log transform, imputation, dan leakage tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah transformasi log untuk rentang aktivitas beberapa orde. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah melakukan scaling menggunakan seluruh dataset sebelum split. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membangun pipeline praproses yang di-fit hanya pada data latih. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap standardization, normalization, log transform, imputation, dan leakage harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus transformasi log untuk rentang aktivitas beberapa orde?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 5.6 praproses data: scaling, transformasi, dan missing value perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti standardization dan hubungan operasionalnya dengan normalization. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu fitur dengan rentang berbeda memerlukan scaling; distribusi menceng dan missing value harus ditangani tanpa kebocoran informasi dari data uji, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan standardization mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar log transform. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah normalization, log transform, dan imputation harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus transformasi log untuk rentang aktivitas beberapa orde, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan standardization–leakage mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan standardization, normalization, log transform, imputation, leakage sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 5.6 praproses data: scaling, transformasi, dan missing value perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk transformasi log untuk rentang aktivitas beberapa orde, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi standardization, atau konversi normalization—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada

kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan log transform sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa melakukan scaling menggunakan seluruh dataset sebelum split; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 5.6 praproses data: scaling, transformasi, dan missing value dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk transformasi log untuk rentang aktivitas beberapa orde, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan

eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.6 praproses data: scaling, transformasi, dan missing value, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk transformasi log untuk rentang aktivitas beberapa orde,

mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai standardization–leakage, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus transformasi log untuk rentang aktivitas beberapa orde dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membangun pipeline praproses yang di-fit hanya pada data latih sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling

efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.6 praproses data: scaling, transformasi, dan missing value sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang standardization akan memengaruhi cara normalization dihitung, cara log transform dibandingkan, dan cara imputation dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada transformasi log untuk rentang aktivitas beberapa orde', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi standardization, normalization, log transform, imputation, leakage dengan data eksperimen,

literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
standardization	Bagaimana standardization didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
leakage	Bagaimana leakage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.7 Korelasi, Seleksi Fitur, dan Reduksi Dimensi

Seleksi fitur dan pca mengurangi redundansi, meningkatkan interpretabilitas, serta mengendalikan rasio sampel terhadap fitur. Pokok bahasan ini menempatkan correlation filter, variance threshold, PCA, feature selection, dan regularization sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu

menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah correlation filter, variance threshold, PCA, feature selection, dan regularization tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah lima komponen utama menjelaskan sebagian besar variasi deskriptor. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih fitur berdasarkan hubungan dengan seluruh endpoint termasuk data uji. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan model dengan dan tanpa seleksi fitur. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap correlation filter, variance threshold, PCA, feature selection, dan regularization harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
correlation filter	Peran correlation filter pada konteks Korelasi, Seleksi Fitur, dan Reduksi Dimensi.
variance threshold	Peran variance threshold pada konteks Korelasi, Seleksi Fitur, dan Reduksi Dimensi.
PCA	Peran PCA pada konteks Korelasi, Seleksi Fitur, dan Reduksi Dimensi.
feature selection	Peran feature selection pada konteks Korelasi, Seleksi Fitur, dan Reduksi Dimensi.
regularization	Peran regularization pada konteks Korelasi, Seleksi Fitur, dan Reduksi Dimensi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus lima komponen utama menjelaskan sebagian besar variasi deskriptor?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 5.7 korelasi, seleksi fitur, dan reduksi dimensi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti correlation filter dan hubungan operasionalnya dengan variance threshold. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu seleksi fitur dan PCA mengurangi redundansi, meningkatkan interpretabilitas, serta mengendalikan rasio sampel terhadap fitur, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan correlation filter mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar PCA. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah variance threshold, PCA, dan feature selection harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus lima komponen utama menjelaskan sebagian besar variasi deskriptor, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan correlation filter–regularization mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan correlation filter, variance threshold, PCA, feature selection,

regularization sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 5.7 korelasi, seleksi fitur, dan reduksi dimensi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk lima komponen utama menjelaskan sebagian besar variasi deskriptor, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi correlation filter, atau konversi variance threshold—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar

senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan PCA sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih fitur berdasarkan hubungan dengan seluruh endpoint termasuk data uji; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 5.7 korelasi, seleksi fitur, dan reduksi dimensi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk lima komponen utama menjelaskan sebagian besar variasi deskriptor, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode.

Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.7 korelasi, seleksi fitur, dan reduksi dimensi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk lima komponen utama menjelaskan sebagian besar variasi deskriptor, mahasiswa perlu menjelaskan unit independensi:

apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai correlation filter–regularization, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus lima komponen utama menjelaskan sebagian besar variasi deskriptor dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas membandingkan model dengan dan tanpa seleksi fitur sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.7 korelasi, seleksi fitur, dan reduksi dimensi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang correlation filter akan memengaruhi cara variance threshold dihitung, cara PCA dibandingkan, dan cara feature selection dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada lima komponen utama menjelaskan sebagian besar variasi deskriptor', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi correlation filter, variance threshold, PCA, feature selection, regularization dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus

entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
correlation filter	Bagaimana correlation filter didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
regularization	Bagaimana regularization memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

5.8 Visualisasi Ruang Kimia dan Batas Interpretasi

Proyeksi pca, t-sne, umap, som, dan peta berbasis jarak membantu eksplorasi, tetapi representasi dua dimensi dapat mendistorsi jarak global. Pokok bahasan ini menempatkan PCA, t-SNE, UMAP, SOM, dan projection sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah PCA, t-SNE, UMAP, SOM, dan projection tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau

pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah memetakan koleksi aktif dan inaktif untuk melihat overlap. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyimpulkan klaster kimia hanya dari satu visualisasi tanpa parameter dan validasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat dua proyeksi dan membandingkan struktur lokal-global. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap PCA, t-SNE, UMAP, SOM, dan projection harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus memetakan koleksi aktif dan inaktif untuk melihat overlap?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 5.8 visualisasi ruang kimia dan batas interpretasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti PCA dan hubungan operasionalnya dengan t-SNE. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu proyeksi PCA, t-SNE, UMAP, SOM, dan peta berbasis jarak membantu eksplorasi, tetapi representasi dua dimensi dapat mendistorsi jarak global, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan PCA mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar UMAP. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah t-SNE, UMAP, dan SOM harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus memetakan koleksi aktif dan inaktif untuk melihat overlap, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan PCA–projection mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan PCA, t-SNE, UMAP, SOM, projection sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 5.8 visualisasi ruang kimia dan batas interpretasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk memetakan koleksi aktif dan inaktif untuk melihat overlap, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi PCA, atau konversi t-SNE—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan UMAP sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyimpulkan klaster kimia hanya dari satu visualisasi tanpa parameter dan validasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV

dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 5.8 visualisasi ruang kimia dan batas interpretasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk memetakan koleksi aktif dan inaktif untuk melihat overlap, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 5.8 visualisasi ruang kimia dan batas interpretasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk memetakan koleksi aktif dan inaktif untuk melihat overlap, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai PCA–projection, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus memetakan koleksi aktif dan inaktif untuk melihat overlap dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membuat dua proyeksi dan membandingkan struktur lokal-global sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 5.8 visualisasi ruang kimia dan batas interpretasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang PCA akan memengaruhi cara t-SNE dihitung, cara UMAP dibandingkan, dan cara SOM dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada memetakan koleksi aktif dan inaktif untuk melihat overlap', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi PCA, t-SNE, UMAP, SOM, projection dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCA	Bagaimana PCA didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
projection	Bagaimana projection memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$z = (x - \mu) / \sigma$$

Standardisasi z

$$t_k = X\, p_k, \text{ dengan } p_k \text{ sebagai loading komponen utama ke-k.}$$

PCA

Praktikum 5 — Deskriptor, PCA, dan Peta Ruang Kimia

Menghitung deskriptor RDKit, scaling, PCA, dan interpretasi loading tanpa data leakage.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from rdkit import Chem
from rdkit.Chem import Descriptors
import pandas as pd

def calc(smiles):
    mol = Chem.MolFromSmiles(smiles)
    if mol is None:
        raise ValueError(f"SMILES tidak valid: {smiles}")
```

```
    return {
        "MW": Descriptors.MolWt(mol),
        "logP": Descriptors.MolLogP(mol),
        "TPSA": Descriptors.TPSA(mol),
        "HBD": Descriptors.NumHDonors(mol),
        "HBA": Descriptors.NumHAcceptors(mol),
    }

df = pd.DataFrame([calc(s) for s in ["CCO", "CC(=O)O", "c1ccccc1"]])
print(df)
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Deskriptor dan Ruang Kimia harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

Jelaskan dengan contoh hubungan antara 0D dan vektor fitur pada konteks berat molekul, jumlah donor ikatan hidrogen, indeks topologi, dan volume.

Identifikasi minimal tiga risiko jika peneliti menggunakan ribuan deskriptor tanpa interpretasi atau kontrol kebocoran.

Jelaskan dengan contoh hubungan antara molecular weight dan HBA pada konteks profil sederhana untuk menilai kelarutan dan permeabilitas awal.

Identifikasi minimal tiga risiko jika peneliti menganggap nilai komputasi identik dengan pengukuran eksperimen.

Jelaskan dengan contoh hubungan antara Wiener dan BCUT pada konteks membedakan isomer dengan konektivitas berbeda namun rumus molekul sama.

Identifikasi minimal tiga risiko jika peneliti menumpuk deskriptor sangat berkorelasi yang menurunkan stabilitas model.

Jelaskan dengan contoh hubungan antara MACCS dan bit collision pada konteks ECFP radius 2 menangkap lingkungan lokal sampai dua ikatan.

Identifikasi minimal tiga risiko jika peneliti menginterpretasikan bit hashed sebagai fitur tunggal yang pasti.

Jelaskan dengan contoh hubungan antara shape dan field descriptor pada konteks deskriptor permukaan bermuatan untuk prediksi permeabilitas.

Identifikasi minimal tiga risiko jika peneliti menghitung satu konformer tanpa menguji sensitivitas.

Jelaskan dengan contoh hubungan antara standardization dan leakage pada konteks transformasi log untuk rentang aktivitas beberapa orde.

Identifikasi minimal tiga risiko jika peneliti melakukan scaling menggunakan seluruh dataset sebelum split.

Jelaskan dengan contoh hubungan antara correlation filter dan regularization pada konteks lima komponen utama menjelaskan sebagian besar variasi deskriptor.

Identifikasi minimal tiga risiko jika peneliti memilih fitur berdasarkan hubungan dengan seluruh endpoint termasuk data uji.

Jelaskan dengan contoh hubungan antara PCA dan projection pada konteks memetakan koleksi aktif dan inaktif untuk melihat overlap.

Identifikasi minimal tiga risiko jika peneliti menyimpulkan klaster kimia hanya dari satu visualisasi tanpa parameter dan validasi.

Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.

Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.

Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.

Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap

kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 6. KEMIRIPAN, KEANEKARAGAMAN, KLASTERING, DAN SELEKSI SENYAWA

Orientasi Bab

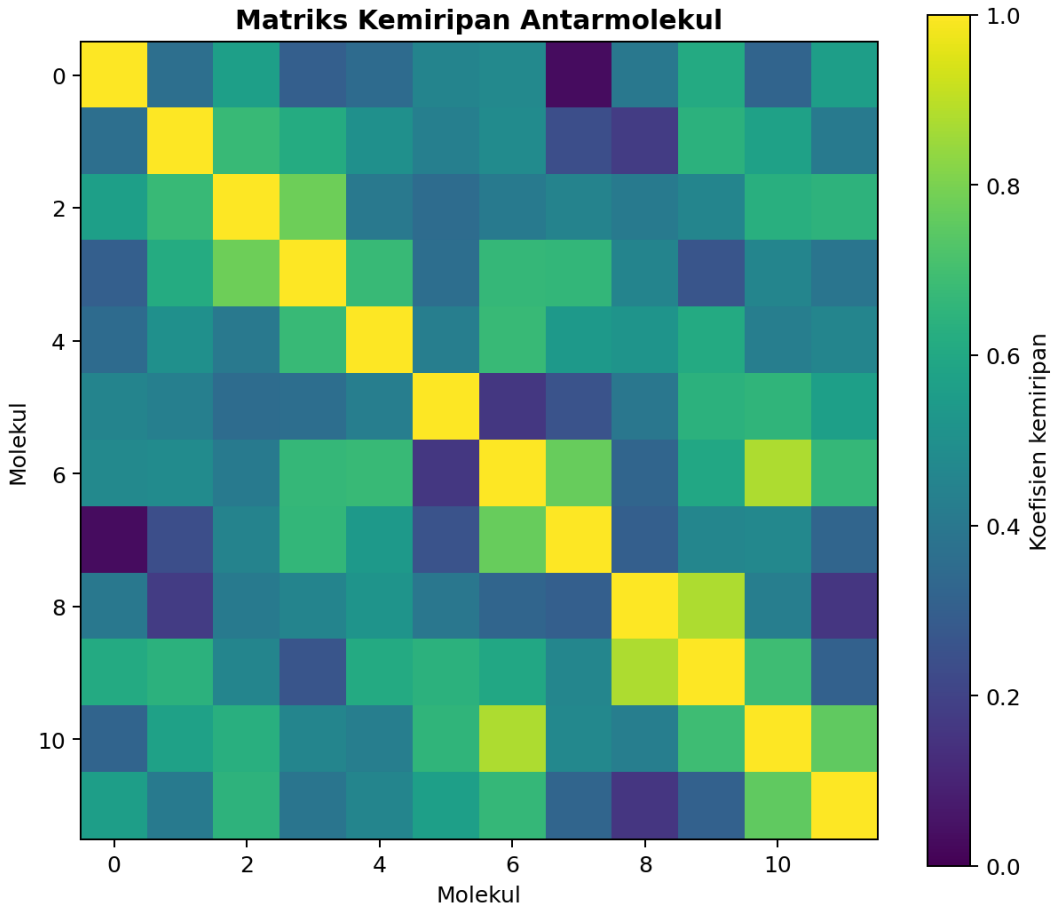
Bab ini menyusun fondasi konseptual dan praktis mengenai kemiripan dan keanekaragaman. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti kemiripan dan keanekaragaman dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

activity cliff, atom typing, availability, binary vector, cell-based, cluster split, cosine, cost, coverage, curse of dimensionality, DBSCAN, decision, dendrogram, Dice, Euclidean, external validation, hierarchical, k-medoids, linkage, locality, Mahalanobis, Manhattan, max-min, MCS, multi-criteria, quality, random split, reduced graph, representative set, SAR, scaffold, scaffold hopping, scaffold split, scaling, similar property principle, sphere exclusion, Tanimoto, time split, timeout, Tversky.



Gambar 6.1. Matriks kemiripan yang dapat menjadi dasar clustering dan pemilihan analog.

6.1 Prinsip Sifat Serupa dan Activity Cliff

Prinsip bahwa struktur serupa cenderung memiliki sifat serupa adalah heuristik penting, tetapi activity cliff menunjukkan perubahan kecil dapat menghasilkan perubahan aktivitas besar. Pokok bahasan ini menempatkan similar property principle, activity cliff, scaffold hopping, locality, dan SAR sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah similar property principle, activity cliff, scaffold hopping, locality, dan SAR tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi

operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah substitusi kecil yang mengubah orientasi interaksi kunci. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap kemiripan tinggi selalu menjamin aktivitas sama. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mencari pasangan serupa dengan selisih aktivitas besar. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap similar property principle, activity cliff, scaffold hopping, locality, dan SAR harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
similar property principle	Peran similar property principle pada konteks Prinsip Sifat Serupa dan Activity Cliff.
activity cliff	Peran activity cliff pada konteks Prinsip Sifat Serupa dan Activity Cliff.
scaffold hopping	Peran scaffold hopping pada konteks Prinsip Sifat Serupa dan Activity Cliff.
locality	Peran locality pada konteks Prinsip Sifat Serupa dan Activity Cliff.
SAR	Peran SAR pada konteks Prinsip Sifat Serupa dan Activity Cliff.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus substitusi kecil yang mengubah orientasi interaksi kunci?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 6.1 prinsip sifat serupa dan activity cliff perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti similar property principle dan hubungan operasionalnya dengan activity cliff. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas,

identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu prinsip bahwa struktur serupa cenderung memiliki sifat serupa adalah heuristik penting, tetapi activity cliff menunjukkan perubahan kecil dapat menghasilkan perubahan aktivitas besar, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan similar property principle mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar scaffold hopping. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah activity cliff, scaffold hopping, dan locality harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus substitusi kecil yang mengubah orientasi interaksi kunci, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan similar property principle–SAR mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan similar property principle, activity cliff, scaffold hopping, locality, SAR sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 6.1 prinsip sifat serupa dan activity cliff perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk substitusi kecil yang mengubah orientasi interaksi kunci, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi similar property principle, atau konversi activity cliff—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan scaffold hopping sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus

cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap kemiripan tinggi selalu menjamin aktivitas sama; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 6.1 prinsip sifat serupa dan activity cliff dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus

mempunyai arti kimia yang dapat diinterpretasikan. Untuk substitusi kecil yang mengubah orientasi interaksi kunci, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.1 prinsip sifat serupa dan activity cliff, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk substitusi kecil yang mengubah orientasi interaksi kunci, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai similar property principle–SAR, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus substitusi kecil yang mengubah orientasi interaksi kunci dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas mencari pasangan serupa dengan selisih aktivitas besar sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.1 prinsip sifat serupa dan activity cliff sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang similar property principle akan memengaruhi cara activity cliff dihitung, cara scaffold hopping dibandingkan, dan cara locality dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada substitusi kecil yang mengubah orientasi interaksi kunci', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi similar property principle, activity cliff, scaffold hopping, locality, SAR dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas

kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similar property principle	Bagaimana similar property principle didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.2 Koefisien Kemiripan untuk Fingerprint

Tanimoto, dice, cosine, tversky, dan koefisien lain memberikan penekanan berbeda pada fitur bersama dan fitur unik. Pokok bahasan ini menempatkan Tanimoto, Dice, cosine, Tversky, dan binary vector sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah Tanimoto, Dice, cosine, Tversky, dan binary vector tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah dua fingerprint dengan 16 bit bersama dari total fitur berbeda. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah membandingkan skor dari fingerprint yang berbeda seolah berada pada skala identik. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menghitung manual Tanimoto dan Dice. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Tanimoto, Dice, cosine, Tversky, dan binary vector harus selalu dihubungkan dengan bab lain. Representasi

memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus dua fingerprint dengan 16 bit bersama dari total fitur berbeda?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 6.2 koefisien kemiripan untuk fingerprint perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Tanimoto dan hubungan operasionalnya dengan Dice. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu Tanimoto, Dice, cosine, Tversky, dan koefisien lain memberikan penekanan berbeda pada fitur bersama dan fitur unik, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Tanimoto mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar cosine. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah Dice, cosine, dan Tversky harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus dua fingerprint dengan 16 bit bersama dari total fitur berbeda, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan

administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Tanimoto–binary vector mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan Tanimoto, Dice, cosine, Tversky, binary vector sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 6.2 koefisien kemiripan untuk fingerprint perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki

pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk dua fingerprint dengan 16 bit bersama dari total fitur berbeda, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Tanimoto, atau konversi Dice—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan cosine sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa membandingkan skor dari fingerprint yang berbeda seolah berada pada skala identik; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 6.2 koefisien kemiripan untuk fingerprint dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk dua fingerprint dengan 16 bit bersama dari total fitur berbeda, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline

cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.2

koefisien kemiripan untuk fingerprint, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk dua fingerprint dengan 16 bit bersama dari total fitur berbeda, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Tanimoto–binary vector, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus dua fingerprint dengan 16 bit bersama dari total fitur berbeda dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas menghitung manual Tanimoto dan Dice sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi

mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.2 koefisien kemiripan untuk fingerprint sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Tanimoto akan memengaruhi cara Dice dihitung, cara cosine dibandingkan, dan cara Tversky dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada dua fingerprint dengan 16 bit bersama dari total fitur berbeda', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu

dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Tanimoto, Dice, cosine, Tversky, binary vector dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Tanimoto	Bagaimana Tanimoto didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
binary vector	Bagaimana binary vector memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.3 Jarak untuk Deskriptor Kontinu

Euclidean, manhattan, mahalanobis, cosine distance, dan metrik lain memerlukan perhatian pada scaling, korelasi, dan dimensi. Pokok bahasan ini menempatkan Euclidean, Manhattan, Mahalanobis, scaling, dan curse of dimensionality sebagai

komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah Euclidean, Manhattan, Mahalanobis, scaling, dan curse of dimensionality tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan Euclidean pada fitur tidak diskalakan. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada

menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan ranking tetangga sebelum dan sesudah scaling. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Euclidean, Manhattan, Mahalanobis, scaling, dan curse of dimensionality harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
Euclidean	Peran Euclidean pada konteks Jarak untuk Deskriptor Kontinu.
Manhattan	Peran Manhattan pada konteks Jarak untuk Deskriptor Kontinu.
Mahalanobis	Peran Mahalanobis pada konteks Jarak untuk Deskriptor Kontinu.
scaling	Peran scaling pada konteks Jarak untuk Deskriptor Kontinu.
curse of dimensionality	Peran curse of dimensionality pada konteks Jarak untuk Deskriptor Kontinu.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 6.3 jarak untuk deskriptor kontinu perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Euclidean dan hubungan operasionalnya dengan Manhattan. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu Euclidean, Manhattan, Mahalanobis, cosine distance, dan metrik lain memerlukan perhatian pada scaling, korelasi, dan dimensi, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Euclidean mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar Mahalanobis. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah Manhattan, Mahalanobis, dan scaling harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Euclidean–curse of dimensionality mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan Euclidean, Manhattan, Mahalanobis, scaling, curse of

dimensionality sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 6.3 jarak untuk deskriptor kontinu perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Euclidean, atau konversi Manhattan—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan

atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan Mahalanobis sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan Euclidean pada fitur tidak diskalakan; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 6.3 jarak untuk deskriptor kontinu dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang

relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.3 jarak untuk deskriptor kontinu, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya.

Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Euclidean–curse of dimensionality, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi dapat dirancang sebagai keputusan bertahap. Tahap pertama

menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas membandingkan ranking tetangga sebelum dan sesudah scaling sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.3 jarak untuk deskriptor kontinu sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Euclidean akan memengaruhi cara Manhattan dihitung, cara Mahalanobis dibandingkan, dan cara scaling dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Euclidean, Manhattan, Mahalanobis, scaling, curse of dimensionality dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Euclidean	Bagaimana Euclidean didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
curse of dimensionality	Bagaimana curse of dimensionality memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.4 Kemiripan Substruktur, MCS, dan Reduced Graph

Maximum common substructure dan reduced graph memberikan interpretasi scaffold, tetapi biaya komputasi dan definisi kesetaraan kimia harus dikendalikan. Pokok bahasan ini menempatkan MCS, scaffold, reduced graph, atom typing, dan timeout sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah MCS, scaffold, reduced graph, atom typing, dan timeout tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna

molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menemukan inti bersama analog inhibitor. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

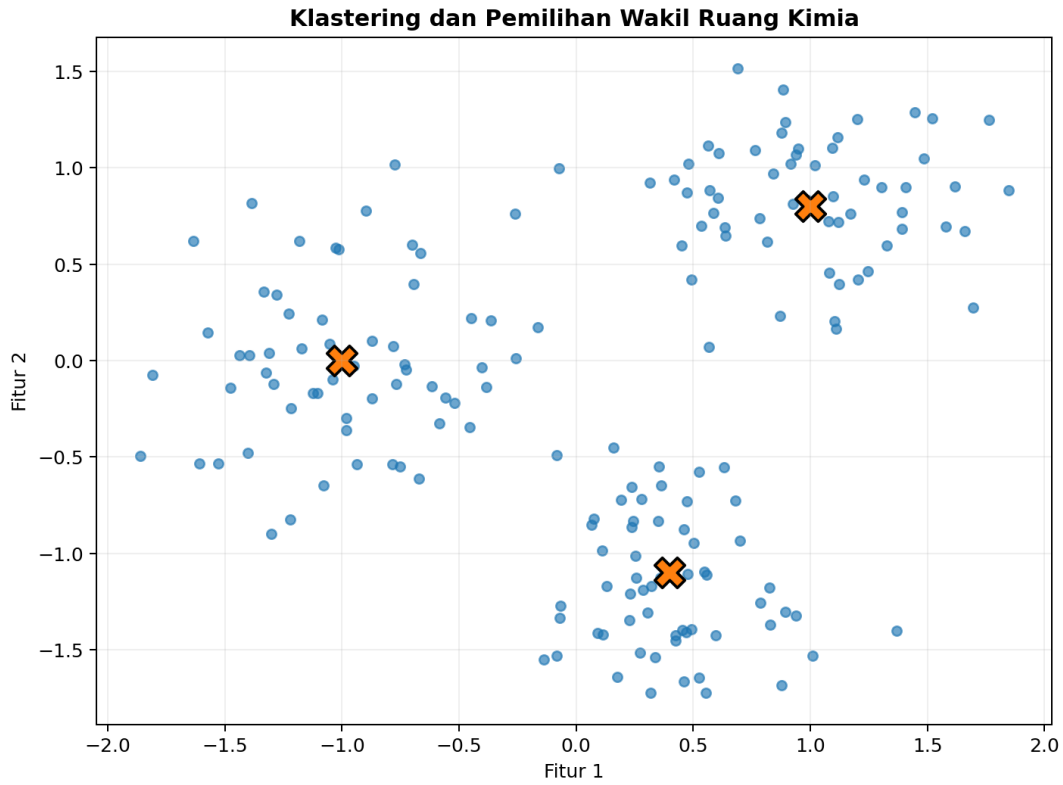
Kesalahan yang sering muncul adalah membiarkan parameter MCS terlalu longgar sehingga hasil tidak bermakna. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengidentifikasi MCS untuk tiga analog sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap MCS, scaffold, reduced graph, atom typing, dan timeout harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 6.2. Klastering ruang kimia dan pemilihan pusat atau medoid sebagai wakil.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menemukan inti bersama analog inhibitor?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 6.4 kemiripan substruktur, mcs, dan reduced graph perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti MCS dan hubungan operasionalnya dengan scaffold. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu maximum common substructure dan reduced graph memberikan interpretasi scaffold, tetapi biaya komputasi dan definisi kesetaraan kimia harus dikendalikan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan MCS mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar reduced graph. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah scaffold, reduced graph, dan atom typing harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menemukan inti bersama analog inhibitor, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan MCS–timeout mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan MCS, scaffold, reduced graph, atom typing, timeout sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 6.4 kemiripan substruktur, mcs, dan reduced graph perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menemukan inti bersama analog inhibitor, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi MCS, atau konversi scaffold—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika

satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan reduced graph sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa membiarkan parameter MCS terlalu longgar sehingga hasil tidak bermakna; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 6.4 kemiripan substruktur, mcs, dan reduced graph dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menemukan inti bersama analog inhibitor, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan,

stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.4 kemiripan substruktur, mcs, dan reduced graph, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menemukan inti bersama analog inhibitor, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum

semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai MCS–timeout, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus menemukan inti bersama analog inhibitor dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan

kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas mengidentifikasi MCS untuk tiga analog sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.4 kemiripan substruktur, mcs, dan reduced graph sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang MCS akan memengaruhi cara scaffold dihitung, cara reduced graph dibandingkan, dan cara atom typing dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menemukan inti bersama analog inhibitor', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi MCS, scaffold, reduced graph, atom typing, timeout dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MCS	Bagaimana MCS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
timeout	Bagaimana timeout memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.5 Klastering Hierarkis dan Nonhierarkis

Agglomerative, divisive, k-means, k-medoids, dbscan, dan metode berbasis graf menghasilkan struktur kelompok yang bergantung pada metrik dan parameter. Pokok bahasan ini menempatkan hierarchical, k-medoids, DBSCAN, linkage, dan dendrogram sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah hierarchical, k-medoids, DBSCAN, linkage, dan dendrogram tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi.

Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah Butina clustering pada fingerprint untuk koleksi senyawa. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih jumlah klaster hanya karena visualisasi tampak rapi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan dua metode klastering pada dataset sama. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap hierarchical, k-medoids, DBSCAN, linkage, dan dendrogram harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
hierarchical	Peran hierarchical pada konteks Klastering Hierarkis dan Nonhierarkis.
k-medoids	Peran k-medoids pada konteks Klastering Hierarkis dan Nonhierarkis.
DBSCAN	Peran DBSCAN pada konteks Klastering Hierarkis dan Nonhierarkis.
linkage	Peran linkage pada konteks Klastering Hierarkis dan Nonhierarkis.
dendrogram	Peran dendrogram pada konteks Klastering Hierarkis dan Nonhierarkis.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus Butina clustering pada fingerprint untuk koleksi senyawa?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam analisis yang lebih rinci, unit analisis pada 6.5 klastering hierarkis dan nonhierarkis perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti hierarchical dan hubungan operasionalnya dengan k-medoids. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama,

aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif kimia, representasi selalu bersifat selektif. Fokus bagian ini, yaitu agglomerative, divisive, k-means, k-medoids, DBSCAN, dan metode berbasis graf menghasilkan struktur kelompok yang bergantung pada metrik dan parameter, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan hierarchical mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar DBSCAN. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif komputasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah k-medoids, DBSCAN, dan linkage harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus Butina clustering pada fingerprint untuk koleksi senyawa, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada tahap verifikasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan hierarchical–dendrogram mengubah jumlah entitas unik serta interpretasi hasil.

Untuk pembelajaran berbasis proyek, peta konsep pada bagian ini dapat dibangun dengan menempatkan hierarchical, k-medoids, DBSCAN, linkage, dendrogram sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam skenario industri, sumber data untuk 6.5 klastering hierarkis dan nonhierarkis perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk Butina clustering pada fingerprint untuk koleksi senyawa, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam riset akademik, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi hierarchical, atau konversi k-medoids—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pertama, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan DBSCAN sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga

kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Selanjutnya, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih jumlah klaster hanya karena visualisasi tampak rapi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada sisi lain, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam analisis yang lebih rinci, algoritme pada 6.5 klastering hierarkis dan nonhierarkis dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan

keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk Butina clustering pada fingerprint untuk koleksi senyawa, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif kimia, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif komputasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada tahap verifikasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Untuk pembelajaran berbasis proyek, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam skenario industri, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.5 klastering hierarkis dan nonhierarkis, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam riset akademik, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk Butina clustering pada fingerprint untuk koleksi senyawa, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pertama, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Selanjutnya, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai hierarchical–dendrogram, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada sisi lain, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam analisis yang lebih rinci, studi kasus Butina clustering pada fingerprint untuk koleksi senyawa dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif kimia, aktivitas membandingkan dua metode klastering pada dataset sama sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif komputasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada tahap verifikasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Untuk pembelajaran berbasis proyek, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.5 klastering hierarkis dan nonhierarkis sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang hierarchical akan memengaruhi cara k-medoids dihitung, cara DBSCAN dibandingkan, dan cara linkage dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada Butina clustering pada fingerprint untuk koleksi senyawa', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi hierarchical, k-medoids, DBSCAN, linkage, dendrogram dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
hierarchical	Bagaimana hierarchical didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
dendrogram	Bagaimana dendrogram memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.6 Pemilihan Set Beragam

Max-min, sphere exclusion, d-optimal, cell-based, dan strategi dissimilarity bertujuan memilih subset informatif yang menutupi ruang kimia. Pokok bahasan ini menempatkan max-min, sphere exclusion, cell-based, coverage, dan representative set sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah max-min, sphere exclusion, cell-based, coverage, dan representative set tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah memilih 100 senyawa dari 10.000 untuk uji eksperimen. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengejar keragaman maksimum sambil mengabaikan kelayakan eksperimen. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menerapkan greedy max-min dan mengevaluasi cakupan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap max-min, sphere exclusion, cell-based, coverage, dan representative set harus selalu dihubungkan dengan bab lain.

Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus memilih 100 senyawa dari 10.000 untuk uji eksperimen?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 6.6 pemilihan set beragam perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti max-min dan hubungan operasionalnya dengan sphere exclusion. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu max-min, sphere exclusion, D-optimal, cell-based, dan strategi dissimilarity bertujuan memilih subset informatif yang menutupi ruang kimia, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan max-min mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar cell-based. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah sphere exclusion, cell-based, dan coverage harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus memilih 100 senyawa dari 10.000 untuk uji eksperimen, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan

hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan max-min–representative set mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan max-min, sphere exclusion, cell-based, coverage, representative set sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 6.6 pemilihan set beragam perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau

menggunakan identifier yang berubah. Untuk memilih 100 senyawa dari 10.000 untuk uji eksperimen, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi max-min, atau konversi sphere exclusion—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan cell-based sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengejar keragaman maksimum sambil mengabaikan kelayakan eksperimen; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 6.6 pemilihan set beragam dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk memilih 100 senyawa dari 10.000 untuk uji eksperimen, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap

akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.6 pemilihan set

beragam, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk memilih 100 senyawa dari 10.000 untuk uji eksperimen, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai max-min–representative set, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus memilih 100 senyawa dari 10.000 untuk uji eksperimen dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas menerapkan greedy max-min dan mengevaluasi cakupan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.6 pemilihan set beragam sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang max-min akan memengaruhi cara sphere exclusion dihitung, cara cell-based dibandingkan, dan cara coverage dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada memilih 100 senyawa dari 10.000 untuk uji eksperimen', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu

dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi max-min, sphere exclusion, cell-based, coverage, representative set dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
max-min	Bagaimana max-min didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
representative set	Bagaimana representative set memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.7 Desain Set Latih, Validasi, dan Uji

Pemisahan acak sering terlalu optimistis; scaffold split, time split, cluster split, dan external set menguji generalisasi yang berbeda. Pokok bahasan ini menempatkan

random split, scaffold split, time split, cluster split, dan external validation sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah random split, scaffold split, time split, cluster split, dan external validation tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menghindari analog dekat tersebar antara train dan test. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyebut test set eksternal padahal berasal dari sumber dan periode sama. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh,

menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat dua skema split dan membandingkan performa. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap random split, scaffold split, time split, cluster split, dan external validation harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
random split	Peran random split pada konteks Desain Set Latih, Validasi, dan Uji.
scaffold split	Peran scaffold split pada konteks Desain Set Latih, Validasi, dan Uji.
time split	Peran time split pada konteks Desain Set Latih, Validasi, dan Uji.
cluster split	Peran cluster split pada konteks Desain Set Latih, Validasi, dan Uji.
external validation	Peran external validation pada konteks Desain Set Latih, Validasi, dan Uji.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menghindari analog dekat tersebar antara train dan test?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 6.7 desain set latih, validasi, dan uji perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti random split dan hubungan operasionalnya dengan scaffold split. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu pemisahan acak sering terlalu optimistis; scaffold split, time split, cluster split, dan external set menguji generalisasi yang berbeda, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan random split mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar time split. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah scaffold split, time split, dan cluster split harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menghindari analog dekat tersebar antara train dan test, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan random split–external validation mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan random split, scaffold split, time split, cluster split, external validation sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 6.7 desain set latih, validasi, dan uji perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menghindari analog dekat tersebar antara train dan test, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi random split, atau konversi scaffold split—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok

tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan time split sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyebut test set eksternal padahal berasal dari sumber dan periode sama; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 6.7 desain set latih, validasi, dan uji dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menghindari analog dekat tersebar antara train dan test, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.7 desain set latih, validasi, dan uji, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menghindari analog dekat tersebar antara train dan test, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-

benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai random split–external validation, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus menghindari analog dekat tersebar antara train dan test dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas membuat dua skema split dan membandingkan performa sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.7 desain set latih, validasi, dan uji sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang random split akan memengaruhi cara scaffold split dihitung, cara time split dibandingkan, dan cara cluster split dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menghindari analog dekat tersebar antara train dan test', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi random split, scaffold split, time split, cluster split, external validation dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus

entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
random split	Bagaimana random split didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
external validation	Bagaimana external validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

6.8 Keanekaragaman Multi-Kriteria dan Keputusan Praktis

Seleksi nyata menggabungkan keragaman struktur, sifat, biaya, ketersediaan, stabilitas, keselamatan, dan tujuan biologis. Pokok bahasan ini menempatkan multi-criteria, cost, availability, quality, dan decision sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah multi-criteria, cost, availability, quality, dan decision tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam

aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah memilih panel senyawa untuk praktikum dengan keterbatasan anggaran. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengubah semua kriteria menjadi satu skor tanpa analisis sensitivitas. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak

menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun matriks keputusan dan menjelaskan trade-off. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap multi-criteria, cost, availability, quality, dan decision harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus memilih panel senyawa untuk praktikum dengan keterbatasan anggaran?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 6.8 keanekaragaman multi-kriteria dan keputusan praktis perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti multi-criteria dan hubungan operasionalnya dengan cost. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu seleksi nyata menggabungkan keragaman struktur, sifat, biaya, ketersediaan, stabilitas, keselamatan, dan tujuan biologis, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan multi-criteria mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar availability. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana

yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah cost, availability, dan quality harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus memilih panel senyawa untuk praktikum dengan keterbatasan anggaran, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan multi-criteria–decision mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan multi-criteria, cost, availability, quality, decision sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 6.8 keanekaragaman multi-kriteria dan keputusan praktis perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk memilih panel senyawa untuk praktikum dengan keterbatasan anggaran, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi multi-criteria, atau konversi cost—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan availability sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengubah semua kriteria menjadi

satu skor tanpa analisis sensitivitas; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 6.8 keanekaragaman multi-kriteria dan keputusan praktis dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk memilih panel senyawa untuk praktikum dengan keterbatasan anggaran, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis

sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 6.8 keanekaragaman multi-kriteria dan keputusan praktis, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk memilih panel senyawa untuk praktikum dengan keterbatasan anggaran, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai multi-criteria–decision, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus memilih panel senyawa untuk praktikum dengan keterbatasan anggaran dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas menyusun matriks keputusan dan menjelaskan trade-off sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 6.8 keanekaragaman multi-kriteria dan keputusan praktis sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang multi-criteria akan memengaruhi cara cost dihitung, cara availability dibandingkan, dan

cara quality dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada memilih panel senyawa untuk praktikum dengan keterbatasan anggaran', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi multi-criteria, cost, availability, quality, decision dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multi-criteria	Bagaimana multi-criteria didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision	Bagaimana decision memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$T = c / (a + b - c)$$

Tanimoto

$$D = 2c / (a + b)$$

Dice

$$d = 1 - T$$

Dissimilarity

Praktikum 6 — Kemiripan, Klastering, dan Seleksi Beragam

Menghitung fingerprint Morgan, matriks Tanimoto, Butina clustering, dan memilih wakil setiap klaster.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.

Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from rdkit import Chem, DataStructs
from rdkit.Chem import AllChem

a = Chem.MolFromSmiles("CC(=O)Oc1ccccc1C(=O)O")
b = Chem.MolFromSmiles("CC(=O)NC1=CC=C(C=C1)O")
fa = AllChem.GetMorganGenerator(radius=2, fpSize=2048).GetFingerprint(a)
fb = AllChem.GetMorganGenerator(radius=2, fpSize=2048).GetFingerprint(b)
print(DataStructs.TanimotoSimilarity(fa, fb))
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Kemiripan dan Keanekaragaman harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara similar property principle dan SAR pada konteks substitusi kecil yang mengubah orientasi interaksi kunci.
2. Identifikasi minimal tiga risiko jika peneliti menganggap kemiripan tinggi selalu menjamin aktivitas sama.
3. Jelaskan dengan contoh hubungan antara Tanimoto dan binary vector pada konteks dua fingerprint dengan 16 bit bersama dari total fitur berbeda.
4. Identifikasi minimal tiga risiko jika peneliti membandingkan skor dari fingerprint yang berbeda seolah berada pada skala identik.
5. Jelaskan dengan contoh hubungan antara Euclidean dan curse of dimensionality pada konteks jarak molekul berdasarkan MW, logP, TPSA setelah standardisasi.
6. Identifikasi minimal tiga risiko jika peneliti menggunakan Euclidean pada fitur tidak diskalakan.
7. Jelaskan dengan contoh hubungan antara MCS dan timeout pada konteks menemukan inti bersama analog inhibitor.
8. Identifikasi minimal tiga risiko jika peneliti membiarkan parameter MCS terlalu longgar sehingga hasil tidak bermakna.
9. Jelaskan dengan contoh hubungan antara hierarchical dan dendrogram pada konteks Butina clustering pada fingerprint untuk koleksi senyawa.
10. Identifikasi minimal tiga risiko jika peneliti memilih jumlah klaster hanya karena visualisasi tampak rapi.
11. Jelaskan dengan contoh hubungan antara max-min dan representative set pada konteks memilih 100 senyawa dari 10.000 untuk uji eksperimen.
12. Identifikasi minimal tiga risiko jika peneliti mengejar keragaman maksimum sambil mengabaikan kelayakan eksperimen.
13. Jelaskan dengan contoh hubungan antara random split dan external validation pada konteks menghindari analog dekat tersebar antara train dan test.
14. Identifikasi minimal tiga risiko jika peneliti menyebut test set eksternal padahal berasal dari sumber dan periode sama.
15. Jelaskan dengan contoh hubungan antara multi-criteria dan decision pada konteks memilih panel senyawa untuk praktikum dengan keterbatasan anggaran.
16. Identifikasi minimal tiga risiko jika peneliti mengubah semua kriteria menjadi satu skor tanpa analisis sensitivitas.
17. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.

18. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
19. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
20. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 7. QSAR, QSPR, PEMBELAJARAN MESIN, DAN VALIDASI MODEL

Orientasi Bab

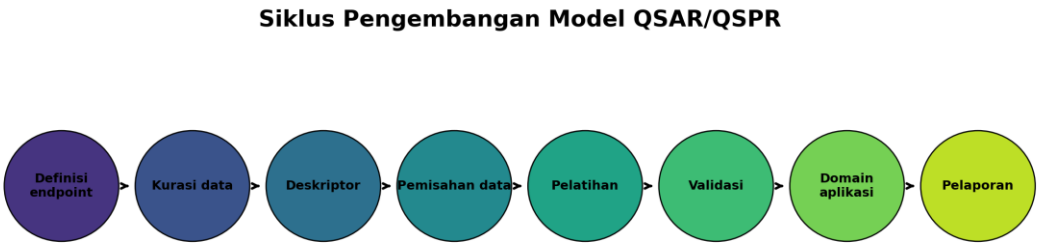
Bab ini menyusun fondasi konseptual dan praktis mengenai qsar dan pembelajaran mesin. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti qsar dan pembelajaran mesin dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

applicability domain, assay, atom attribution, boosting, calibration, coefficient, confidence interval, conformal prediction, cross-validation, decision context, decision tree, elastic net, embedding, endpoint, external set, GNN, hyperparameter, interpretability, interpretation, k-NN, kernel, lasso, leverage, logistic regression, MAE, margin, MCC, message passing, MLR, MMP, model card, multicollinearity, nested CV, neural network, OOD, PCR, PLS, pretraining, probability, R2, random forest, residual, ridge, RMSE, SHAP, SVM, task definition, uncertainty, unit, variance.



Gambar 7.1. Siklus pengembangan model QSAR/QSPR yang memisahkan data, fitur, validasi, dan pelaporan.

7.1 Perumusan Endpoint dan Desain Studi Prediktif

Model dimulai dari definisi endpoint yang konsisten, unit yang jelas, protokol yang sebanding, dan tujuan penggunaan yang terukur. Pokok bahasan ini menempatkan endpoint, assay, unit, task definition, dan decision context sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah endpoint, assay, unit, task definition, dan decision context tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah memprediksi logS pada pH dan suhu tertentu. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara

komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mencampur IC50 dari assay berbeda tanpa harmonisasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis definisi endpoint operasional dan kriteria inklusi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap endpoint, assay, unit, task definition, dan decision context harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
endpoint	Peran endpoint pada konteks Perumusan Endpoint dan Desain Studi Prediktif.
assay	Peran assay pada konteks Perumusan Endpoint dan Desain Studi Prediktif.
unit	Peran unit pada konteks Perumusan Endpoint dan Desain Studi Prediktif.
task definition	Peran task definition pada konteks Perumusan Endpoint dan Desain Studi Prediktif.
decision context	Peran decision context pada konteks Perumusan Endpoint dan Desain Studi Prediktif.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus memprediksi logS pada pH dan suhu tertentu?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 7.1 perumusan endpoint dan desain studi prediktif perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti endpoint dan hubungan operasionalnya dengan assay. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model dimulai dari definisi endpoint yang konsisten, unit yang jelas, protokol yang sebanding, dan tujuan penggunaan yang terukur, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan endpoint mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar unit. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah assay, unit, dan task definition harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang

berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus memprediksi logS pada pH dan suhu tertentu, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan endpoint–decision context mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan endpoint, assay, unit, task definition, decision context sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 7.1 perumusan endpoint dan desain studi prediktif perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk memprediksi logS pada pH dan suhu tertentu, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi endpoint, atau konversi assay—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan unit sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mencampur IC50 dari assay berbeda tanpa harmonisasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks

mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 7.1 perumusan endpoint dan desain studi prediktif dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk memprediksi logS pada pH dan suhu tertentu, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian

subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.1 perumusan endpoint dan desain studi prediktif, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk memprediksi logS pada pH dan suhu tertentu, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai endpoint–decision context, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus memprediksi logS pada pH dan suhu tertentu dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas menulis definisi endpoint operasional dan kriteria inklusi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama,

mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.1 perumusan endpoint dan desain studi prediktif sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang endpoint akan memengaruhi cara assay dihitung, cara unit dibandingkan, dan cara task definition dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada memprediksi logS pada pH dan suhu tertentu', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi endpoint, assay, unit, task definition, decision context dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
endpoint	Bagaimana endpoint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decision context	Bagaimana decision context memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.2 Regresi Linear, MLR, dan Dasar QSAR Klasik

Regresi linear menghubungkan respons dengan deskriptor melalui koefisien yang dapat ditafsirkan, tetapi memerlukan asumsi, pemeriksaan residual, dan kontrol multikolinearitas. Pokok bahasan ini menempatkan MLR, coefficient, residual, multicollinearity, dan interpretation sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam kerangka pengambilan keputusan, istilah MLR, coefficient, residual, multicollinearity, dan interpretation tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah hubungan log aktivitas dengan hidrofobisitas dan efek elektronik. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menafsirkan korelasi sebagai sebab-akibat. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak

valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membangun model dua-deskriptor dan memeriksa residual. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap MLR, coefficient, residual, multicollinearity, dan interpretation harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus hubungan log aktivitas dengan hidrofobisitas dan efek elektronik?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 7.2 regresi linear, mlr, dan dasar qsar klasik perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti MLR dan hubungan operasionalnya dengan coefficient. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu regresi linear menghubungkan respons dengan deskriptor melalui koefisien yang dapat ditafsirkan, tetapi memerlukan asumsi, pemeriksaan residual, dan kontrol multikolinearitas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan MLR mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar residual. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah coefficient, residual, dan multicollinearity harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus hubungan log aktivitas dengan hidrofobisitas dan efek elektronik, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan MLR–interpretation mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan MLR, coefficient, residual, multicollinearity, interpretation sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 7.2 regresi linear, mlr, dan dasar qsar klasik perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk hubungan log aktivitas dengan hidrofobisitas dan efek elektronik, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi MLR, atau konversi coefficient—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika

satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan residual sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menafsirkan korelasi sebagai sebab-akibat; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 7.2 regresi linear, mlr, dan dasar qsar klasik dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk hubungan log aktivitas dengan hidrofobisitas dan efek elektronik, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan,

stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.2 regresi linear, mlr, dan dasar qsar klasik, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk hubungan log aktivitas dengan hidrofobisitas dan efek elektronik, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya.

Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai MLR–interpretation, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus hubungan log aktivitas dengan hidrofobisitas dan efek elektronik dapat dirancang sebagai keputusan bertahap. Tahap pertama

menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas membangun model dua-deskriptor dan memeriksa residual sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.2 regresi linear, mlr, dan dasar qsar klasik sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang MLR akan memengaruhi cara coefficient dihitung, cara residual dibandingkan, dan cara multicollinearity dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada hubungan log aktivitas dengan hidrofobisitas dan efek elektronik', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi MLR, coefficient, residual, multicollinearity, interpretation dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
MLR	Bagaimana MLR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
interpretation	Bagaimana interpretation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.3 PCR, PLS, Regularisasi, dan Seleksi Variabel

Pcr dan pls mengatasi banyak fitur berkorelasi, sedangkan ridge, lasso, dan elastic net mengendalikan kompleksitas melalui penalti. Pokok bahasan ini menempatkan PCR, PLS, ridge, lasso, dan elastic net sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah PCR, PLS, ridge, lasso, dan elastic net tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah model spektrum atau descriptor matrix dengan p lebih besar dari n. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih komponen berdasarkan performa test set. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menentukan hiperparameter melalui cross-validation tersarang. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan

demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap PCR, PLS, ridge, lasso, dan elastic net harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
PCR	Peran PCR pada konteks PCR, PLS, Regularisasi, dan Seleksi Variabel.
PLS	Peran PLS pada konteks PCR, PLS, Regularisasi, dan Seleksi Variabel.
ridge	Peran ridge pada konteks PCR, PLS, Regularisasi, dan Seleksi Variabel.
lasso	Peran lasso pada konteks PCR, PLS, Regularisasi, dan Seleksi Variabel.
elastic net	Peran elastic net pada konteks PCR, PLS, Regularisasi, dan Seleksi Variabel.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus model spektrum atau descriptor matrix dengan p lebih besar dari n?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 7.3 pcr, pls, regularisasi, dan seleksi variabel perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti PCR dan hubungan operasionalnya dengan PLS. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu PCR dan PLS mengatasi banyak fitur berkorelasi, sedangkan ridge, lasso, dan elastic net mengendalikan kompleksitas melalui penalti, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang

diabaikan. Representasi yang menekankan PCR mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar ridge. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah PLS, ridge, dan lasso harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus model spektrum atau descriptor matrix dengan p lebih besar dari n, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan PCR–elastic net mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan PCR, PLS, ridge, lasso, elastic net sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 7.3 pcr, pls, regularisasi, dan seleksi variabel perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk model spektrum atau descriptor matrix dengan p lebih besar dari n, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi PCR, atau konversi PLS—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan ridge sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier,

struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih komponen berdasarkan performa test set; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 7.3 pcr, pls, regularisasi, dan seleksi variabel dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk model spektrum atau descriptor matrix dengan p lebih besar dari n, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.3 pcr, pls, regularisasi, dan seleksi variabel, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk model spektrum atau descriptor matrix dengan p lebih besar dari n, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai

PCR–elastic net, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus model spektrum atau descriptor matrix dengan p lebih besar dari n dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas menentukan hiperparameter melalui cross-validation tersarang sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data

terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.3 pcr, pls, regularisasi, dan seleksi variabel sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang PCR akan memengaruhi cara PLS dihitung, cara ridge dibandingkan, dan cara lasso dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada model spektrum atau descriptor matrix dengan p lebih besar dari n', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi PCR, PLS, ridge, lasso, elastic net dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
PCR	Bagaimana PCR didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
elastic net	Bagaimana elastic net memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.4 Klasifikasi: Logistik, Pohon, Random Forest, dan Boosting

Metode klasifikasi menghasilkan probabilitas atau kelas aktif-inaktif dan perlu dievaluasi dengan metrik yang sesuai ketidakseimbangan. Pokok bahasan ini menempatkan logistic regression, decision tree, random forest, boosting, dan probability sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah logistic regression, decision tree, random forest, boosting, dan probability tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah klasifikasi toksik-nontoksik dengan kelas toksik jarang. Contoh perlu dibahas pada tiga

tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

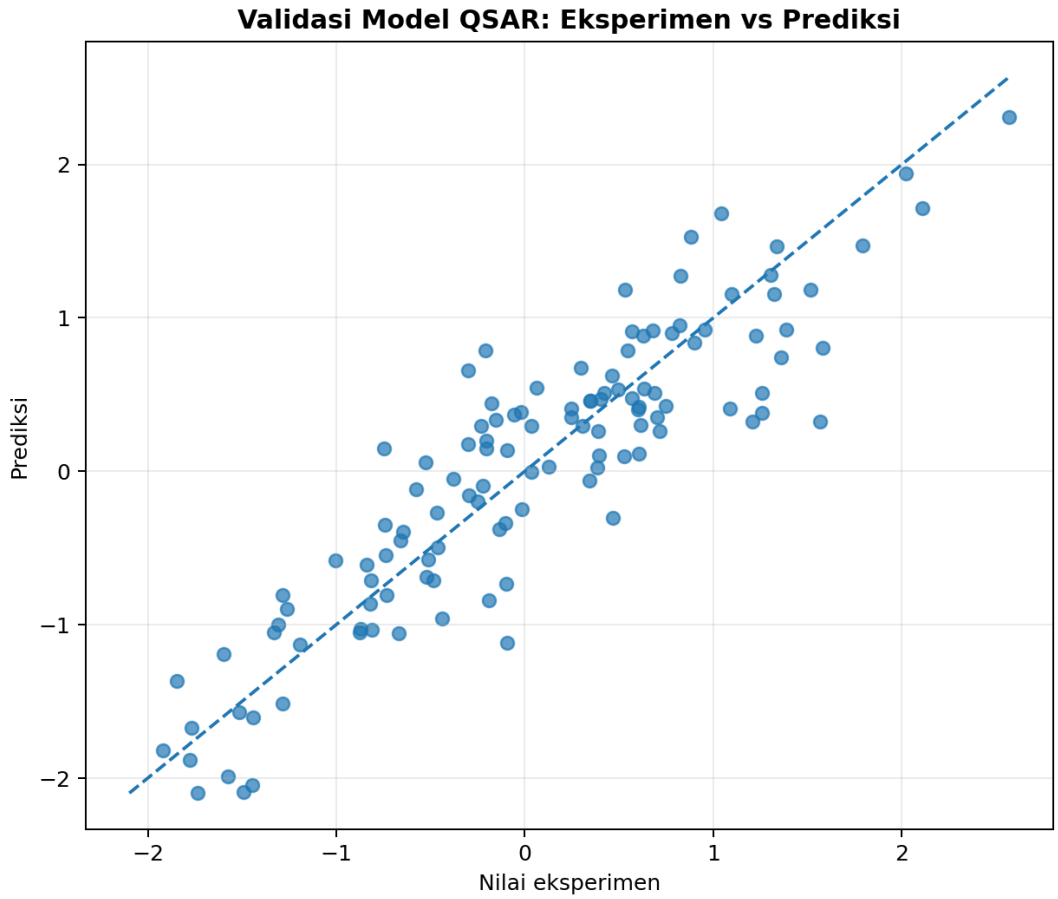
Kesalahan yang sering muncul adalah mengandalkan akurasi pada data sangat tidak seimbang. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan ROC-AUC dan PR-AUC. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap logistic regression, decision tree, random forest, boosting, dan probability harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 7.2. Plot eksperimen versus prediksi sebagai salah satu diagnostik model regresi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus klasifikasi toksik-nontoksik dengan kelas toksik jarang?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 7.4 klasifikasi: logistik, pohon, random forest, dan boosting perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti logistic regression dan hubungan operasionalnya dengan decision tree. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu metode klasifikasi menghasilkan probabilitas atau kelas aktif-inaktif dan perlu dievaluasi dengan metrik yang sesuai ketidakseimbangan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan logistic regression mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar random forest. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah decision tree, random forest, dan boosting harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus klasifikasi toksik-nontoksik dengan kelas toksik jarang, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan logistic regression–probability mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan logistic regression, decision tree, random forest, boosting, probability sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 7.4 klasifikasi: logistik, pohon, random forest, dan boosting perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk klasifikasi toksik-nontoksik dengan kelas toksik jarang, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi logistic regression, atau konversi decision tree—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan random forest sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai

objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengandalkan akurasi pada data sangat tidak seimbang; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 7.4 klasifikasi: logistik, pohon, random forest, dan boosting dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas;

transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk klasifikasi toksik-nontoksik dengan kelas toksik jarang, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.4 klasifikasi: logistik, pohon, random forest, dan boosting, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk klasifikasi toksik-nontoksik dengan kelas toksik jarang, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya

dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai logistic regression–probability, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus klasifikasi toksik-nontoksik dengan kelas toksik jarang dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap

menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas membandingkan ROC-AUC dan PR-AUC sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.4 klasifikasi: logistik, pohon, random forest, dan boosting sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang logistic regression akan memengaruhi cara decision tree dihitung, cara random forest dibandingkan, dan cara boosting dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada klasifikasi toksik-nontoksik dengan kelas toksik jarang', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi logistic regression, decision tree, random forest, boosting, probability dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi

yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
logistic regression	Bagaimana logistic regression didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
probability	Bagaimana probability memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.5 SVM, Kernel, dan Metode Berbasis Tetangga

Svm memaksimalkan margin dan kernel memetakan kesamaan; k-nn menggunakan lokalitas, tetapi keduanya peka pada skala, parameter, dan distribusi ruang kimia. Pokok bahasan ini menempatkan SVM, kernel, margin, k-NN, dan hyperparameter sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah SVM, kernel, margin, k-NN, dan hyperparameter tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi,

penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah kernel Tanimoto untuk fingerprint biner. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyetel parameter menggunakan test set. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah melakukan pencarian hiperparameter dalam cross-validation. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap SVM, kernel, margin, k-NN, dan hyperparameter harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
SVM	Peran SVM pada konteks SVM, Kernel, dan Metode Berbasis Tetangga.
kernel	Peran kernel pada konteks SVM, Kernel, dan Metode Berbasis Tetangga.
margin	Peran margin pada konteks SVM, Kernel, dan Metode Berbasis Tetangga.
k-NN	Peran k-NN pada konteks SVM, Kernel, dan Metode Berbasis Tetangga.
hyperparameter	Peran hyperparameter pada konteks SVM, Kernel, dan Metode Berbasis Tetangga.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus kernel Tanimoto untuk fingerprint biner?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 7.5 svm, kernel, dan metode berbasis tetangga perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti SVM dan hubungan operasionalnya dengan kernel. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu SVM memaksimalkan margin dan kernel memetakan kesamaan; k-NN menggunakan lokalitas, tetapi keduanya peka pada skala, parameter, dan distribusi ruang kimia, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan SVM mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar margin.

Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah kernel, margin, dan k-NN harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus kernel Tanimoto untuk fingerprint biner, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan SVM–hyperparameter mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan SVM, kernel, margin, k-NN, hyperparameter sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 7.5 svm, kernel, dan metode berbasis tetangga perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk kernel Tanimoto untuk fingerprint biner, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi SVM, atau konversi kernel—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan margin sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur

kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyetel parameter menggunakan test set; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 7.5 svm, kernel, dan metode berbasis tetangga dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk kernel Tanimoto untuk fingerprint biner, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria

konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.5 svm, kernel, dan metode berbasis tetangga, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk kernel Tanimoto untuk fingerprint biner, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai SVM–hyperparameter, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus kernel Tanimoto untuk fingerprint biner dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas melakukan pencarian hiperparameter dalam cross-validation sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.5 svm, kernel, dan metode berbasis tetangga sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang SVM akan memengaruhi cara kernel dihitung, cara margin dibandingkan, dan cara k-NN

dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada kernel Tanimoto untuk fingerprint biner', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi SVM, kernel, margin, k-NN, hyperparameter dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
SVM	Bagaimana SVM didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
hyperparameter	Bagaimana hyperparameter memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.6 Jaringan Saraf, GNN, dan Representasi Terpelajar

Model neural dapat belajar representasi dari graf, string, atau geometri 3d, tetapi membutuhkan data, regularisasi, interpretasi, dan pembandingan baseline. Pokok bahasan ini menempatkan neural network, message passing, GNN, embedding, dan pretraining sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah neural network, message passing, GNN, embedding, dan pretraining tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah message passing mengagregasi informasi tetangga atom. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap model lebih dalam selalu lebih baik. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan GNN dengan fingerprint plus random forest. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap neural network, message passing, GNN, embedding, dan pretraining harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus message passing mengagregasi informasi tetangga atom?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 7.6 jaringan saraf, gnn, dan representasi terpelajar perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti neural network dan hubungan operasionalnya dengan message passing. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model neural dapat belajar representasi dari graf, string, atau geometri 3D, tetapi membutuhkan data, regularisasi, interpretasi, dan pembandingan baseline, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan neural network mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar GNN. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah message passing, GNN, dan embedding harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus message passing mengagregasi informasi tetangga atom, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan neural network–pretraining mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan neural network, message passing, GNN, embedding, pretraining sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 7.6 jaringan saraf, gnn, dan representasi terpelajar perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk message passing mengagregasi informasi tetangga atom, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi neural network, atau konversi message passing—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada

kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan GNN sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap model lebih dalam selalu lebih baik; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 7.6 jaringan saraf, gnn, dan representasi terpelajar dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk message passing mengagregasi informasi tetangga atom, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.6 jaringan saraf, gnn, dan representasi terpelajar, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk message passing mengagregasi informasi tetangga atom, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris

benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai neural network–pretraining, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus message passing mengagregasi informasi tetangga atom dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas membandingkan GNN dengan fingerprint plus random forest sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.6 jaringan saraf, gnn, dan representasi terpelajar sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang neural network akan memengaruhi cara message passing dihitung, cara GNN dibandingkan, dan cara embedding dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada message passing mengagregasi informasi tetangga atom', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi neural network, message passing, GNN, embedding, pretraining dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas,

membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
neural network	Bagaimana neural network didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
pretraining	Bagaimana pretraining memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.7 Validasi Internal, Eksternal, dan Nested Cross-Validation

Validasi harus memisahkan pemilihan model dari estimasi performa dan melaporkan variasi antar-split, bukan hanya satu angka terbaik. Pokok bahasan ini menempatkan cross-validation, nested CV, external set, confidence interval, dan variance sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah cross-validation, nested CV, external set, confidence interval, dan variance tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat

mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah outer loop mengukur performa, inner loop memilih hiperparameter. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah data leakage melalui standardisasi atau seleksi fitur sebelum split. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menggambar diagram nested cross-validation. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap cross-validation, nested CV, external set, confidence interval, dan variance harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
cross-validation	Peran cross-validation pada konteks Validasi Internal, Eksternal, dan Nested Cross-Validation.
nested CV	Peran nested CV pada konteks Validasi Internal, Eksternal, dan Nested Cross-Validation.
external set	Peran external set pada konteks Validasi Internal, Eksternal, dan Nested Cross-Validation.
confidence interval	Peran confidence interval pada konteks Validasi Internal, Eksternal, dan Nested Cross-Validation.
variance	Peran variance pada konteks Validasi Internal, Eksternal, dan Nested Cross-Validation.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus outer loop mengukur performa, inner loop memilih hiperparameter?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 7.7 validasi internal, eksternal, dan nested cross-validation perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti cross-validation dan hubungan operasionalnya dengan nested CV. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan

objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu validasi harus memisahkan pemilihan model dari estimasi performa dan melaporkan variasi antar-split, bukan hanya satu angka terbaik, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan cross-validation mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar external set. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah nested CV, external set, dan confidence interval harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus outer loop mengukur performa, inner loop memilih hiperparameter, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan cross-validation–variance mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan cross-validation, nested CV, external set, confidence interval, variance sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang

menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 7.7 validasi internal, eksternal, dan nested cross-validation perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk outer loop mengukur performa, inner loop memilih hiperparameter, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi cross-validation, atau konversi nested CV—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan external set sebelum koreksi struktur dapat mempertahankan kesalahan

ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa data leakage melalui standardisasi atau seleksi fitur sebelum split; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 7.7 validasi internal, eksternal, dan nested cross-validation dapat dipahami melalui empat komponen: masukan,

transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk outer loop mengukur performa, inner loop memilih hiperparameter, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.7 validasi internal, eksternal, dan nested cross-validation, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk outer loop mengukur performa, inner loop memilih hiperparameter, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya

dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai cross-validation–variance, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus outer loop mengukur performa, inner loop memilih hiperparameter dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen.

Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas menggambar diagram nested cross-validation sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.7 validasi internal, eksternal, dan nested cross-validation sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang cross-validation akan memengaruhi cara nested CV dihitung, cara external set dibandingkan, dan cara confidence interval dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada outer loop mengukur performa, inner loop memilih hiperparameter', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi cross-validation, nested CV, external set, confidence interval, variance dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu

menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
cross-validation	Bagaimana cross-validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
variance	Bagaimana variance memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.8 Metrik Regresi, Klasifikasi, Ranking, dan Kalibrasi

Rmse, mae, r^2, q^2, mcc, f1, auc, enrichment, dan calibration error menjawab pertanyaan berbeda sehingga harus dipilih sesuai penggunaan. Pokok bahasan ini menempatkan RMSE, MAE, R2, MCC, dan calibration sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah RMSE, MAE, R2, MCC, dan calibration tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah model screening memerlukan ranking awal yang baik selain klasifikasi global. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan R^2 tinggi tanpa memeriksa error absolut. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat tabel metrik dan interpretasinya. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak

pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap RMSE, MAE, R2, MCC, dan calibration harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus model screening memerlukan ranking awal yang baik selain klasifikasi global?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 7.8 metrik regresi, klasifikasi, ranking, dan kalibrasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti RMSE dan hubungan operasionalnya dengan MAE. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu RMSE, MAE, R^2, Q^2, MCC, F1, AUC, enrichment, dan calibration error menjawab pertanyaan berbeda sehingga harus dipilih sesuai penggunaan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan RMSE mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar R2. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah MAE, R2, dan MCC harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus model screening memerlukan ranking awal yang baik selain klasifikasi global, satu nilai tanpa konteks dapat

menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan RMSE–calibration mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan RMSE, MAE, R2, MCC, calibration sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 7.8 metrik regresi, klasifikasi, ranking, dan kalibrasi perlu dinilai pada empat lapis: otoritas sumber, cara

pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk model screening memerlukan ranking awal yang baik selain klasifikasi global, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi RMSE, atau konversi MAE—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan R2 sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan R^2 tinggi tanpa memeriksa error absolut; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 7.8 metrik regresi, klasifikasi, ranking, dan kalibrasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk model screening memerlukan ranking awal yang baik selain klasifikasi global, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline

cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.8 metrik

regresi, klasifikasi, ranking, dan kalibrasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk model screening memerlukan ranking awal yang baik selain klasifikasi global, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai RMSE–calibration, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus model screening memerlukan ranking awal yang baik selain klasifikasi global dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas membuat tabel metrik dan interpretasinya sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi

mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.8 metrik regresi, klasifikasi, ranking, dan kalibrasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang RMSE akan memengaruhi cara MAE dihitung, cara R2 dibandingkan, dan cara MCC dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada model screening memerlukan ranking awal yang baik selain klasifikasi global', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan.

Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi RMSE, MAE, R2, MCC, calibration dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
RMSE	Bagaimana RMSE didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
calibration	Bagaimana calibration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.9 Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution

Prediksi paling dapat dipercaya untuk molekul yang cukup mirip dengan data latih; leverage, jarak, ensemble, conformal prediction, dan uncertainty estimation membantu menilai reliabilitas. Pokok bahasan ini menempatkan applicability domain, leverage, uncertainty, OOD, dan conformal prediction sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah applicability domain, leverage, uncertainty, OOD, dan conformal prediction tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menandai kandidat jauh dari ruang kimia latih. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyajikan semua prediksi dengan kepastian sama. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual

pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menambahkan kolom confidence dan alasan penandaan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap applicability domain, leverage, uncertainty, OOD, dan conformal prediction harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
applicability domain	Peran applicability domain pada konteks Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution.
leverage	Peran leverage pada konteks Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution.
uncertainty	Peran uncertainty pada konteks Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution.

Konsep	Makna operasional dalam bagian ini
OOD	Peran OOD pada konteks Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution.
conformal prediction	Peran conformal prediction pada konteks Domain Aplikasi, Ketidakpastian, dan Deteksi Out-of-Distribution.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menandai kandidat jauh dari ruang kimia latih?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 7.9 domain aplikasi, ketidakpastian, dan deteksi out-of-distribution perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti applicability domain dan hubungan operasionalnya dengan leverage. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu prediksi paling dapat dipercaya untuk molekul yang cukup mirip dengan data latih; leverage, jarak, ensemble, conformal prediction, dan uncertainty estimation membantu menilai reliabilitas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan applicability domain mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar uncertainty. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah leverage, uncertainty, dan OOD harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menandai kandidat jauh dari ruang kimia latih, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi,

protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan applicability domain–conformal prediction mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan applicability domain, leverage, uncertainty, OOD, conformal prediction sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 7.9 domain aplikasi, ketidakpastian, dan deteksi out-of-distribution perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi.

Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menandai kandidat jauh dari ruang kimia latih, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi applicability domain, atau konversi leverage—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan uncertainty sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyajikan semua prediksi dengan kepastian sama; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 7.9 domain aplikasi, ketidakpastian, dan deteksi out-of-distribution dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menandai kandidat jauh dari ruang kimia latih, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline

cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.9 domain

aplikasi, ketidakpastian, dan deteksi out-of-distribution, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menandai kandidat jauh dari ruang kimia latih, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai applicability domain–conformal prediction, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus menandai kandidat jauh dari ruang kimia latih dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas menambahkan kolom confidence dan alasan penandaan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.9 domain aplikasi, ketidakpastian, dan deteksi out-of-distribution sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang applicability domain akan memengaruhi cara leverage dihitung, cara uncertainty dibandingkan, dan cara OOD dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menandai kandidat jauh dari ruang kimia latih', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan

mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi applicability domain, leverage, uncertainty, OOD, conformal prediction dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
applicability domain	Bagaimana applicability domain didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
conformal prediction	Bagaimana conformal prediction memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

7.10 Interpretabilitas, Kausalitas, dan Pelaporan Model

Feature importance, shap, atom attribution, matched molecular pairs, dan analisis residual membantu memahami model, tetapi penjelasan prediktif tidak otomatis bersifat kausal. Pokok bahasan ini menempatkan interpretability, SHAP, atom attribution, MMP, dan model card sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah interpretability, SHAP, atom attribution, MMP, dan model card tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menyoroti substruktur yang mendorong prediksi aktivitas. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan penjelasan post-hoc untuk membuktikan mekanisme biologis. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun model card yang menyatakan data, metrik, batasan, dan penggunaan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap interpretability, SHAP, atom attribution, MMP, dan model card harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menyoroti substruktur yang mendorong prediksi aktivitas?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 7.10 interpretabilitas, kausalitas, dan pelaporan model perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti interpretability dan

hubungan operasionalnya dengan SHAP. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu feature importance, SHAP, atom attribution, matched molecular pairs, dan analisis residual membantu memahami model, tetapi penjelasan prediktif tidak otomatis bersifat kausal, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan interpretability mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar atom attribution. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah SHAP, atom attribution, dan MMP harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menyoroti substruktur yang mendorong prediksi aktivitas, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan interpretability–model card mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan interpretability, SHAP, atom attribution, MMP, model card sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian.

Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 7.10 interpretabilitas, kausalitas, dan pelaporan model perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menyoroti substruktur yang mendorong prediksi aktivitas, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi interpretability, atau konversi SHAP—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat

menyebabkan kebocoran; perhitungan atom attribution sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan penjelasan post-hoc untuk membuktikan mekanisme biologis; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 7.10 interpretabilitas, kausalitas, dan pelaporan model dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menyoroti substruktur yang mendorong prediksi aktivitas, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang

relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 7.10 interpretabilitas, kausalitas, dan pelaporan model, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menyoroti substruktur yang mendorong prediksi aktivitas, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter

model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai interpretability–model card, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus menyoroti substruktur yang mendorong prediksi aktivitas dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas menyusun model card yang menyatakan data, metrik, batasan, dan penggunaan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 7.10 interpretabilitas, kausalitas, dan pelaporan model sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang interpretability akan memengaruhi cara SHAP dihitung, cara atom attribution dibandingkan, dan cara MMP dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menyoroti substruktur yang mendorong prediksi aktivitas', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi interpretability, SHAP, atom attribution, MMP, model card dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas,

membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
interpretability	Bagaimana interpretability didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
model card	Bagaimana model card memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Regresi linear

$$\mathrm{RMSE} = \sqrt{[(1/n)\,\Sigma_i\,(y_i - \hat{y}_i)^2]}$$

RMSE

$$R^2 = 1 - \Sigma(y_i - \hat{y}_i)^2 \,/\, \Sigma(y_i - \bar{y})^2$$

R^2

Praktikum 7 — Model QSAR yang Divalidasi

Membangun baseline regresi/klasifikasi, nested cross-validation, metrik, domain aplikasi, dan model card.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from sklearn.model_selection import KFold, cross_val_score
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestRegressor

model = Pipeline([
    ("scale", StandardScaler()),
    ("model", RandomForestRegressor(n_estimators=300, random_state=42))
])
cv = KFold(n_splits=5, shuffle=True, random_state=42)
scores = cross_val_score(model, X, y, cv=cv,
scoring="neg_root_mean_squared_error")
print("RMSE CV:", -scores.mean())
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- QSAR dan Pembelajaran Mesin harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara endpoint dan decision context pada konteks memprediksi logS pada pH dan suhu tertentu.
2. Identifikasi minimal tiga risiko jika peneliti mencampur IC50 dari assay berbeda tanpa harmonisasi.
3. Jelaskan dengan contoh hubungan antara MLR dan interpretation pada konteks hubungan log aktivitas dengan hidrofobisitas dan efek elektronik.
4. Identifikasi minimal tiga risiko jika peneliti menafsirkan korelasi sebagai sebab-akibat.
5. Jelaskan dengan contoh hubungan antara PCR dan elastic net pada konteks model spektrum atau descriptor matrix dengan p lebih besar dari n.
6. Identifikasi minimal tiga risiko jika peneliti memilih komponen berdasarkan performa test set.
7. Jelaskan dengan contoh hubungan antara logistic regression dan probability pada konteks klasifikasi toksik-nontoksik dengan kelas toksik jarang.
8. Identifikasi minimal tiga risiko jika peneliti mengandalkan akurasi pada data sangat tidak seimbang.
9. Jelaskan dengan contoh hubungan antara SVM dan hyperparameter pada konteks kernel Tanimoto untuk fingerprint biner.

10. Identifikasi minimal tiga risiko jika peneliti menyetel parameter menggunakan test set.
11. Jelaskan dengan contoh hubungan antara neural network dan pretraining pada konteks message passing mengagregasi informasi tetangga atom.
12. Identifikasi minimal tiga risiko jika peneliti menganggap model lebih dalam selalu lebih baik.
13. Jelaskan dengan contoh hubungan antara cross-validation dan variance pada konteks outer loop mengukur performa, inner loop memilih hiperparameter.
14. Identifikasi minimal tiga risiko jika peneliti data leakage melalui standardisasi atau seleksi fitur sebelum split.
15. Jelaskan dengan contoh hubungan antara RMSE dan calibration pada konteks model screening memerlukan ranking awal yang baik selain klasifikasi global.
16. Identifikasi minimal tiga risiko jika peneliti menggunakan R^2 tinggi tanpa memeriksa error absolut.
17. Jelaskan dengan contoh hubungan antara applicability domain dan conformal prediction pada konteks menandai kandidat jauh dari ruang kimia latih.
18. Identifikasi minimal tiga risiko jika peneliti menyajikan semua prediksi dengan kepastian sama.
19. Jelaskan dengan contoh hubungan antara interpretability dan model card pada konteks menyoroti substruktur yang mendorong prediksi aktivitas.
20. Identifikasi minimal tiga risiko jika peneliti menggunakan penjelasan post-hoc untuk membuktikan mekanisme biologis.
21. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
22. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
23. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
24. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 8. HIGH-THROUGHPUT SCREENING, PENAPISAN VIRTUAL, DOCKING, DAN ADMET

Orientasi Bab

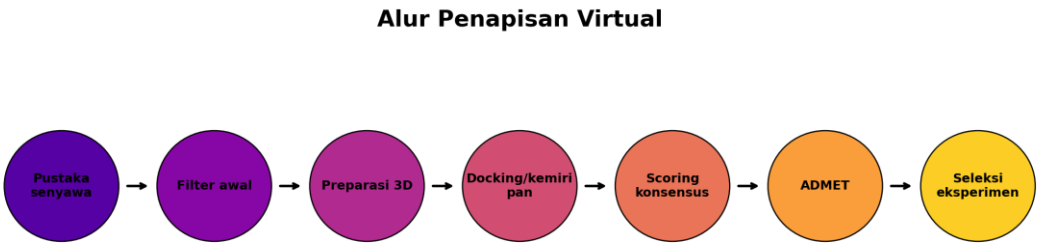
Bab ini menyusun fondasi konseptual dan praktis mengenai penapisan virtual dan admet. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti penapisan virtual dan admet dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

ADMET, artifact, binding site, cofactor, confirmation, consensus, decoy, docking, dose response, enrichment, ensemble, flexibility, grid, heatmap, hit rate, HTS, iteration, Lipinski, metabolism, novelty, PAINS, permeability, pharmacophore, plate effect, pose, prospective validation, protein preparation, protonation, QSAR, ranking, reactive group, replicate, rescoring, sampling, SAR, scoring function, search algorithm, shape, similarity search, solubility, toxicity, triage, Veber, Z-prime.



Gambar 8.1. Penapisan virtual sebagai rangkaian triase dari pustaka hingga kandidat eksperimen.

8.1 Data High-Throughput Screening dan Kontrol Kualitas

Hts menghasilkan banyak pengukuran tetapi juga noise, efek batch, plate edge, agregasi, dan artefak assay sehingga qc merupakan tahap utama. Pokok bahasan ini menempatkan HTS, plate effect, Z-prime, replicate, dan artifact sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah HTS, plate effect, Z-prime, replicate, dan artifact tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah kontrol positif dan negatif untuk menilai kualitas plate. Contoh perlu dibahas pada tiga tingkat.

Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan hit tunggal tanpa konfirmasi orthogonal. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menghitung statistik QC dan menandai plate bermasalah. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap HTS, plate effect, Z-prime, replicate, dan artifact harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
HTS	Peran HTS pada konteks Data High-Throughput Screening dan Kontrol Kualitas.
plate effect	Peran plate effect pada konteks Data High-Throughput Screening dan Kontrol Kualitas.
Z-prime	Peran Z-prime pada konteks Data High-Throughput Screening dan Kontrol Kualitas.
replicate	Peran replicate pada konteks Data High-Throughput Screening dan Kontrol Kualitas.
artifact	Peran artifact pada konteks Data High-Throughput Screening dan Kontrol Kualitas.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus kontrol positif dan negatif untuk menilai kualitas plate?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 8.1 data high-throughput screening dan kontrol kualitas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti HTS dan hubungan operasionalnya dengan plate effect. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu HTS menghasilkan banyak pengukuran tetapi juga noise, efek batch, plate edge, agregasi, dan artefak assay sehingga QC merupakan tahap utama, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan HTS mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar Z-prime. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah plate effect, Z-prime, dan replicate harus diterjemahkan ke aturan yang dapat diuji:

bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus kontrol positif dan negatif untuk menilai kualitas plate, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan HTS–artifact mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan HTS, plate effect, Z-prime, replicate, artifact sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 8.1 data high-throughput screening dan kontrol kualitas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk kontrol positif dan negatif untuk menilai kualitas plate, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi HTS, atau konversi plate effect—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan Z-prime sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan hit tunggal tanpa konfirmasi orthogonal; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan.

Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 8.1 data high-throughput screening dan kontrol kualitas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk kontrol positif dan negatif untuk menilai kualitas plate, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat

meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.1 data high-throughput screening dan kontrol kualitas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk kontrol positif dan negatif untuk menilai kualitas plate, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai HTS–artifact, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus kontrol positif dan negatif untuk menilai kualitas plate dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas menghitung statistik QC dan menandai plate bermasalah sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim

utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.1 data high-throughput screening dan kontrol kualitas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang HTS akan memengaruhi cara plate effect dihitung, cara Z-prime dibandingkan, dan cara replicate dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada kontrol positif dan negatif untuk menilai kualitas plate', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi HTS, plate effect, Z-prime, replicate, artifact dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
HTS	Bagaimana HTS didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
artifact	Bagaimana artifact memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.2 Visualisasi dan Penambangan Hit

Distribusi respons, heatmap plate, kurva konsentrasi-respons, enrichment scaffold, dan analisis substruktur membantu menemukan pola yang tidak tampak dari daftar hit. Pokok bahasan ini menempatkan heatmap, dose response, hit rate, enrichment, dan SAR sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah heatmap, dose response, hit rate, enrichment, dan SAR tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyimpulkan aktivitas spesifik dari pola yang disebabkan interferensi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat dashboard QC sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap heatmap, dose response, hit rate, enrichment, dan SAR harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 8.2 visualisasi dan penambangan hit perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti heatmap dan hubungan operasionalnya dengan dose response. Bila unit analisis berubah di tengah

pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu distribusi respons, heatmap plate, kurva konsentrasi-respons, enrichment scaffold, dan analisis substruktur membantu menemukan pola yang tidak tampak dari daftar hit, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan heatmap mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar hit rate. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah dose response, hit rate, dan enrichment harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan heatmap–SAR mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan heatmap, dose response, hit rate, enrichment, SAR sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman

mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 8.2 visualisasi dan penambangan hit perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi heatmap, atau konversi dose response—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan hit rate sebelum koreksi struktur dapat mempertahankan kesalahan ke

tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyimpulkan aktivitas spesifik dari pola yang disebabkan interferensi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 8.2 visualisasi dan penambangan hit dapat dipahami melalui empat komponen: masukan, transformasi,

parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.2 visualisasi dan penambangan hit, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya

dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai heatmap–SAR, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan

eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas membuat dashboard QC sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.2 visualisasi dan penambangan hit sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang heatmap akan memengaruhi cara dose response dihitung, cara hit rate dibandingkan, dan cara enrichment dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi heatmap, dose response, hit rate, enrichment, SAR dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat

dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
heatmap	Bagaimana heatmap didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
SAR	Bagaimana SAR memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.3 Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan

Filter sifat dan structural alerts berguna untuk triase, tetapi tidak boleh dipakai sebagai hukum mutlak tanpa konteks target dan mekanisme. Pokok bahasan ini menempatkan Lipinski, Veber, PAINS, reactive group, dan triage sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah Lipinski, Veber, PAINS, reactive group, dan triage tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting

dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menandai senyawa reaktif untuk tinjauan manual. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah membuang semua pelanggar aturan walau tujuan bukan obat oral. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat kebijakan filter bertingkat: hard, soft, dan review. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Lipinski, Veber, PAINS, reactive group, dan triage harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
Lipinski	Peran Lipinski pada konteks Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan.
Veber	Peran Veber pada konteks Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan.
PAINS	Peran PAINS pada konteks Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan.
reactive group	Peran reactive group pada konteks Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan.
triage	Peran triage pada konteks Filter Drug-Likeness, PAINS, dan Aturan Kewaspadaan.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menandai senyawa reaktif untuk tinjauan manual?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 8.3 filter drug-likeness, pains, dan aturan kewaspadaan perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Lipinski dan hubungan operasionalnya dengan Veber. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu filter sifat dan structural alerts berguna untuk triase, tetapi tidak boleh dipakai sebagai hukum mutlak tanpa konteks target dan mekanisme, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Lipinski mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar PAINS. Sebaliknya, representasi yang kaya dapat

memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah Veber, PAINS, dan reactive group harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menandai senyawa reaktif untuk tinjauan manual, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Lipinski–triage mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan Lipinski, Veber, PAINS, reactive group, triage sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 8.3 filter drug-likeness, pains, dan aturan kewaspadaan perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menandai senyawa reaktif untuk tinjauan manual, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Lipinski, atau konversi Veber—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan PAINS sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh

pipeline. Pada bagian ini, kesalahan khas berupa membuang semua pelanggar aturan walau tujuan bukan obat oral; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 8.3 filter drug-likeness, pains, dan aturan kewaspadaan dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menandai senyawa reaktif untuk tinjauan manual, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi

mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.3 filter drug-likeness, pains, dan aturan kewaspadaan, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menandai senyawa reaktif untuk tinjauan manual, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Lipinski–triage, serta kemungkinan penyebab kimia. Pola error dapat

mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus menandai senyawa reaktif untuk tinjauan manual dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas membuat kebijakan filter bertingkat: hard, soft, dan review sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook

tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.3 filter drug-likeness, pains, dan aturan kewaspadaan sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Lipinski akan memengaruhi cara Veber dihitung, cara PAINS dibandingkan, dan cara reactive group dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menandai senyawa reaktif untuk tinjauan manual', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Lipinski, Veber, PAINS, reactive group, triage dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Lipinski	Bagaimana Lipinski didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
triage	Bagaimana triage memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.4 Ligand-Based Virtual Screening

Pencarian kemiripan, pharmacophore, shape, qsar, dan machine learning menggunakan informasi ligan yang diketahui untuk memprioritaskan kandidat. Pokok bahasan ini menempatkan similarity search, pharmacophore, shape, QSAR, dan ensemble sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah similarity search, pharmacophore, shape, QSAR, dan ensemble tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mencari analog dan scaffold hopping dari ligan aktif. Contoh perlu dibahas pada tiga tingkat.

Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

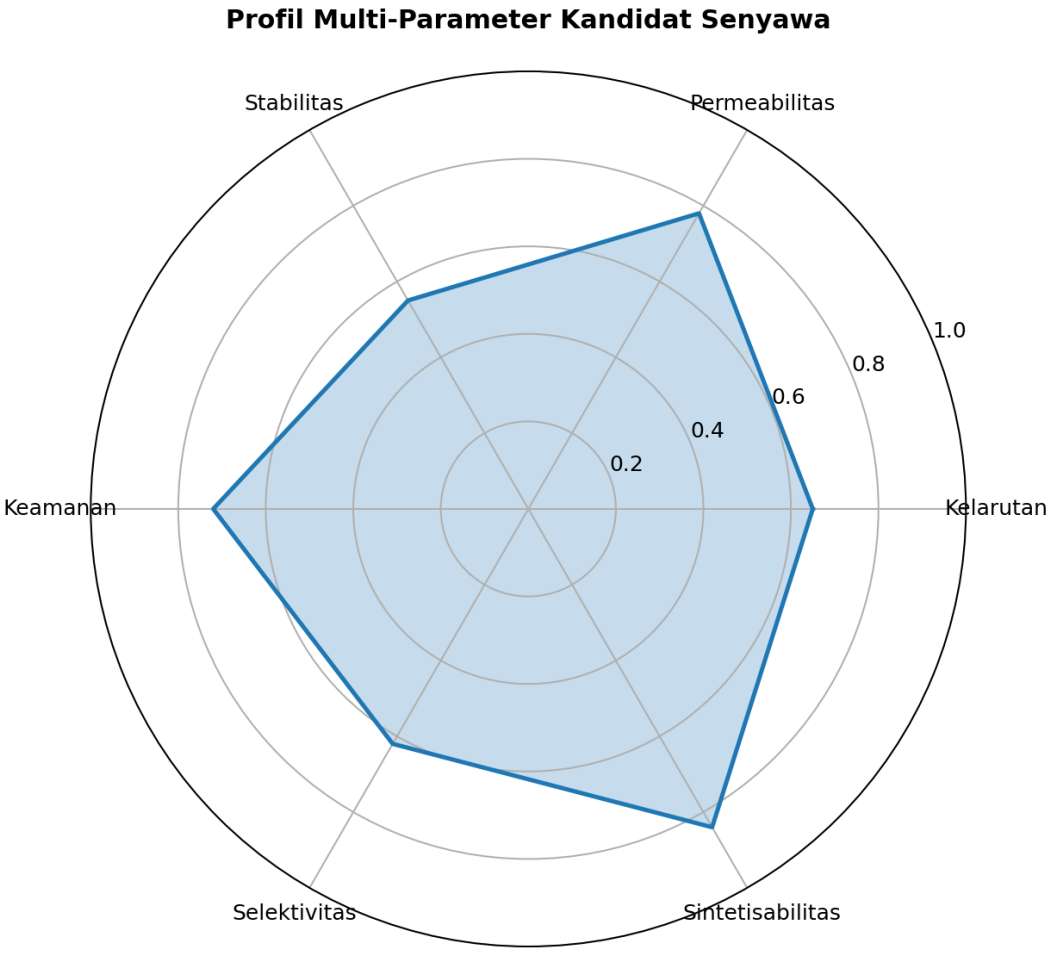
Kesalahan yang sering muncul adalah menggunakan satu metode dan satu referensi tanpa menguji bias. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membangun konsensus beberapa skor ligand-based. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap similarity search, pharmacophore, shape, QSAR, dan ensemble harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 8.2. Profil multi-parameter untuk membandingkan trade-off kandidat.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mencari analog dan scaffold hopping dari ligan aktif?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 8.4 ligand-based virtual screening perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti similarity search dan hubungan operasionalnya dengan pharmacophore. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu pencarian kemiripan, pharmacophore, shape, QSAR, dan machine learning menggunakan informasi ligan yang diketahui untuk memprioritaskan kandidat, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan similarity search mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar shape. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah pharmacophore, shape, dan QSAR harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mencari analog dan scaffold hopping dari ligan aktif, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan similarity search–ensemble mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan similarity search, pharmacophore, shape, QSAR, ensemble sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 8.4 ligand-based virtual screening perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mencari analog dan scaffold hopping dari ligan aktif, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi similarity search, atau konversi pharmacophore—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan shape sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci

sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan satu metode dan satu referensi tanpa menguji bias; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 8.4 ligand-based virtual screening dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus

mempunyai arti kimia yang dapat diinterpretasikan. Untuk mencari analog dan scaffold hopping dari ligan aktif, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.4 ligand-based virtual screening, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mencari analog dan scaffold hopping dari ligan aktif, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai similarity search–ensemble, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus mencari analog dan scaffold hopping dari ligan aktif dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membangun konsensus beberapa skor ligand-based sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.4 ligand-based virtual screening sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang similarity search akan memengaruhi cara pharmacophore dihitung, cara shape dibandingkan, dan cara QSAR dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mencari analog dan scaffold hopping dari ligan aktif', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi similarity search, pharmacophore, shape, QSAR, ensemble dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
similarity search	Bagaimana similarity search didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
ensemble	Bagaimana ensemble memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.5 Structure-Based Virtual Screening dan Preparasi Sistem

Docking memerlukan struktur target, penentuan situs, protonasi, air penting, ion, kofaktor, dan batas kotak yang benar. Pokok bahasan ini menempatkan protein preparation, binding site, protonation, grid, dan cofactor sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Secara konseptual, istilah protein preparation, binding site, protonation, grid, dan cofactor tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah preparasi protease dengan residu katalitik terprotonasi tepat. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mendocking pada struktur tanpa memeriksa ligan ko-kristal dan residu hilang. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis checklist preparasi target. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap protein preparation, binding site, protonation, grid, dan cofactor harus selalu dihubungkan dengan bab lain.

Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
protein preparation	Peran protein preparation pada konteks Structure-Based Virtual Screening dan Preparasi Sistem.
binding site	Peran binding site pada konteks Structure-Based Virtual Screening dan Preparasi Sistem.
protonation	Peran protonation pada konteks Structure-Based Virtual Screening dan Preparasi Sistem.
grid	Peran grid pada konteks Structure-Based Virtual Screening dan Preparasi Sistem.
cofactor	Peran cofactor pada konteks Structure-Based Virtual Screening dan Preparasi Sistem.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus preparasi protease dengan residu katalitik terprotonasi tepat?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 8.5 structure-based virtual screening dan preparasi sistem perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti protein preparation dan hubungan operasionalnya dengan binding site. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu docking memerlukan struktur target, penentuan situs, protonasi, air penting, ion, kofaktor, dan batas kotak yang benar, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan protein preparation mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar protonation. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko

inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah binding site, protonation, dan grid harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus preparasi protease dengan residu katalitik terprotonasi tepat, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan protein preparation–cofactor mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan protein preparation, binding site, protonation, grid, cofactor sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 8.5 structure-based virtual screening dan preparasi sistem perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk preparasi protease dengan residu katalitik terprotonasi tepat, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi protein preparation, atau konversi binding site—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan protonation sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier,

struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mendocking pada struktur tanpa memeriksa ligan ko-kristal dan residu hilang; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 8.5 structure-based virtual screening dan preparasi sistem dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk preparasi protease dengan residu katalitik terprotonasi tepat, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.5 structure-based virtual screening dan preparasi sistem, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk preparasi protease dengan residu katalitik terprotonasi tepat, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada

tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai protein preparation–cofactor, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus preparasi protease dengan residu katalitik terprotonasi tepat dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas menulis checklist preparasi target sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data

mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.5 structure-based virtual screening dan preparasi sistem sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang protein preparation akan memengaruhi cara binding site dihitung, cara protonation dibandingkan, dan cara grid dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada preparasi protease dengan residu katalitik terprotonasi tepat', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi protein preparation, binding site, protonation, grid, cofactor dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
protein preparation	Bagaimana protein preparation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cofactor	Bagaimana cofactor memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.6 Algoritme Docking dan Sampling Pose

Docking menggabungkan eksplorasi orientasi-konformasi dengan fungsi skor; hasil dipengaruhi fleksibilitas ligan, reseptor, dan strategi sampling. Pokok bahasan ini menempatkan pose, sampling, search algorithm, flexibility, dan docking sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah pose, sampling, search algorithm, flexibility, dan docking tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah redocking ligan ko-kristal untuk menguji protokol. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyamakan skor docking dari perangkat berbeda. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengevaluasi keberhasilan pose dengan RMSD dan inspeksi interaksi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap pose, sampling, search algorithm, flexibility, dan docking harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia

adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus redocking ligan ko-kristal untuk menguji protokol?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 8.6 algoritme docking dan sampling pose perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti pose dan hubungan operasionalnya dengan sampling. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu docking menggabungkan eksplorasi orientasi-konformasi dengan fungsi skor; hasil dipengaruhi fleksibilitas ligan, reseptor, dan strategi sampling, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan pose mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar search algorithm. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah sampling, search algorithm, dan flexibility harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus redocking ligan ko-kristal untuk menguji protokol, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi

berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan pose–docking mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan pose, sampling, search algorithm, flexibility, docking sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 8.6 algoritme docking dan sampling pose perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk redocking ligan ko-kristal untuk menguji protokol, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan,

serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi pose, atau konversi sampling—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan search algorithm sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyamakan skor docking dari perangkat berbeda; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 8.6 algoritme docking dan sampling pose dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk redocking ligan ko-kristal untuk menguji protokol, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.6 algoritme docking dan sampling pose, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk redocking ligan ko-kristal untuk menguji protokol, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai pose–docking, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus redocking ligan ko-kristal untuk menguji protokol dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas mengevaluasi keberhasilan pose dengan RMSD dan inspeksi interaksi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.6 algoritme docking dan sampling pose sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang pose akan memengaruhi cara sampling dihitung, cara search algorithm dibandingkan, dan cara flexibility dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada redocking ligan ko-kristal untuk menguji protokol', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi,

sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi pose, sampling, search algorithm, flexibility, docking dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pose	Bagaimana pose didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
docking	Bagaimana docking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.7 Fungsi Skor, Consensus Scoring, dan Rescoring

Fungsi skor empiris, force-field, knowledge-based, dan ml memiliki bias; consensus dan rescoring dapat membantu tetapi membutuhkan validasi. Pokok bahasan ini menempatkan scoring function, consensus, rescoring, ranking, dan decoy sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya,

melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah scoring function, consensus, rescoring, ranking, dan decoy tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menggabungkan skor docking dengan kemiripan interaksi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih kandidat hanya dari skor paling negatif. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan enrichment beberapa strategi scoring. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap scoring function, consensus, rescoring, ranking, dan decoy harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
scoring function	Peran scoring function pada konteks Fungsi Skor, Consensus Scoring, dan Rescoring.
consensus	Peran consensus pada konteks Fungsi Skor, Consensus Scoring, dan Rescoring.
rescoring	Peran rescoring pada konteks Fungsi Skor, Consensus Scoring, dan Rescoring.
ranking	Peran ranking pada konteks Fungsi Skor, Consensus Scoring, dan Rescoring.
decoy	Peran decoy pada konteks Fungsi Skor, Consensus Scoring, dan Rescoring.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menggabungkan skor docking dengan kemiripan interaksi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 8.7 fungsi skor, consensus scoring, dan rescoring perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti scoring function dan hubungan operasionalnya dengan consensus. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu fungsi skor empiris, force-field, knowledge-based, dan ML memiliki bias; consensus dan rescoring dapat membantu tetapi membutuhkan validasi, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan scoring function mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar rescoring. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah consensus, rescoring, dan ranking harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menggabungkan skor docking dengan kemiripan interaksi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan scoring function–decoy mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan scoring function, consensus, rescoring, ranking, decoy sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 8.7 fungsi skor, consensus scoring, dan rescoring perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menggabungkan skor docking dengan kemiripan interaksi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi scoring function, atau konversi consensus—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok

tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan rescoring sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih kandidat hanya dari skor paling negatif; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 8.7 fungsi skor, consensus scoring, dan rescoring dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menggabungkan skor docking dengan kemiripan interaksi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.7 fungsi skor, consensus scoring, dan rescoring, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menggabungkan skor docking dengan kemiripan interaksi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-

benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai scoring function–decoy, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus menggabungkan skor docking dengan kemiripan interaksi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas membandingkan enrichment beberapa strategi scoring sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.7 fungsi skor, consensus scoring, dan rescoring sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang scoring function akan memengaruhi cara consensus dihitung, cara rescoring dibandingkan, dan cara ranking dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menggabungkan skor docking dengan kemiripan interaksi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi scoring function, consensus, rescoring, ranking, decoy dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun

skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scoring function	Bagaimana scoring function didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
decoy	Bagaimana decoy memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.8 Prediksi ADMET dan Toksisitas

Model absorpsi, distribusi, metabolisme, ekskresi, dan toksisitas memakai deskriptor serta data biologis, dengan domain aplikasi dan uncertainty yang harus jelas. Pokok bahasan ini menempatkan ADMET, solubility, permeability, metabolism, dan toxicity sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah ADMET, solubility, permeability, metabolism, dan toxicity tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis.

Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah prediksi BBB, hERG, dan kelarutan untuk prioritas awal. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggantikan uji eksperimen dengan prediksi in silico tanpa validasi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat profil risiko multi-endpoint. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap ADMET, solubility, permeability, metabolism, dan toxicity harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus prediksi BBB, hERG, dan kelarutan untuk prioritas awal?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam analisis yang lebih rinci, unit analisis pada 8.8 prediksi admet dan toksisitas perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti ADMET dan hubungan operasionalnya dengan solubility. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif kimia, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model absorpsi, distribusi, metabolisme, ekskresi, dan toksisitas memakai deskriptor serta data biologis, dengan domain aplikasi dan uncertainty yang harus jelas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan ADMET mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar permeability. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan

informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif komputasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah solubility, permeability, dan metabolism harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus prediksi BBB, hERG, dan kelarutan untuk prioritas awal, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada tahap verifikasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan ADMET–toxicity mengubah jumlah entitas unik serta interpretasi hasil.

Untuk pembelajaran berbasis proyek, peta konsep pada bagian ini dapat dibangun dengan menempatkan ADMET, solubility, permeability, metabolism, toxicity sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam skenario industri, sumber data untuk 8.8 prediksi admet dan toksisitas perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk prediksi BBB, hERG, dan kelarutan untuk prioritas awal, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam riset akademik, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi ADMET, atau konversi solubility—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pertama, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan permeability sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Selanjutnya, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggantikan uji eksperimen dengan prediksi in

silico tanpa validasi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada sisi lain, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam analisis yang lebih rinci, algoritme pada 8.8 prediksi admet dan toksisitas dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk prediksi BBB, hERG, dan kelarutan untuk prioritas awal, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif kimia, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis

sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif komputasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada tahap verifikasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Untuk pembelajaran berbasis proyek, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam skenario industri, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.8 prediksi admet dan toksisitas, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam riset akademik, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk prediksi BBB, hERG, dan kelarutan untuk prioritas awal, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pertama, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Selanjutnya, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai ADMET–toxicity, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada sisi lain, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman

eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam analisis yang lebih rinci, studi kasus prediksi BBB, hERG, dan kelarutan untuk prioritas awal dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif kimia, aktivitas membuat profil risiko multi-endpoint sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif komputasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada tahap verifikasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Untuk pembelajaran berbasis proyek, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.8 prediksi admet dan toksisitas sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang ADMET akan

memengaruhi cara solubility dihitung, cara permeability dibandingkan, dan cara metabolism dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada prediksi BBB, hERG, dan kelarutan untuk prioritas awal', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi ADMET, solubility, permeability, metabolism, toxicity dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ADMET	Bagaimana ADMET didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
toxicity	Bagaimana toxicity memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

8.9 Validasi Prospektif dan Integrasi Eksperimen

Nilai nyata virtual screening diukur melalui hit rate, novelty, potency, selectivity, dan keberhasilan eksperimen prospektif, bukan hanya metrik retrospektif. Pokok bahasan ini menempatkan prospective validation, hit rate, novelty, confirmation, dan iteration sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah prospective validation, hit rate, novelty, confirmation, dan iteration tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mengirim panel terpilih untuk assay konfirmasi dan memperbarui model. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan

berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah melaporkan hanya hit berhasil tanpa seluruh kandidat yang diuji. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah merancang siklus design-make-test-analyze sederhana. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap prospective validation, hit rate, novelty, confirmation, dan iteration harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
prospective validation	Peran prospective validation pada konteks Validasi Prospektif dan Integrasi Eksperimen.

Konsep	Makna operasional dalam bagian ini
hit rate	Peran hit rate pada konteks Validasi Prospektif dan Integrasi Eksperimen.
novelty	Peran novelty pada konteks Validasi Prospektif dan Integrasi Eksperimen.
confirmation	Peran confirmation pada konteks Validasi Prospektif dan Integrasi Eksperimen.
iteration	Peran iteration pada konteks Validasi Prospektif dan Integrasi Eksperimen.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mengirim panel terpilih untuk assay konfirmasi dan memperbarui model?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 8.9 validasi prospektif dan integrasi eksperimen perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti prospective validation dan hubungan operasionalnya dengan hit rate. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu nilai nyata virtual screening diukur melalui hit rate, novelty, potency, selectivity, dan keberhasilan eksperimen prospektif, bukan hanya metrik retrospektif, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan prospective validation mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar novelty. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah hit rate, novelty, dan confirmation harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang

berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mengirim panel terpilih untuk assay konfirmasi dan memperbarui model, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan prospective validation–iteration mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan prospective validation, hit rate, novelty, confirmation, iteration sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 8.9 validasi prospektif dan integrasi eksperimen perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mengirim panel terpilih untuk assay konfirmasi dan memperbarui model, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi prospective validation, atau konversi hit rate—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan novelty sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa melaporkan hanya hit berhasil tanpa seluruh kandidat yang diuji; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks

mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 8.9 validasi prospektif dan integrasi eksperimen dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mengirim panel terpilih untuk assay konfirmasi dan memperbarui model, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian

subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 8.9 validasi prospektif dan integrasi eksperimen, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mengirim panel terpilih untuk assay konfirmasi dan memperbarui model, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai prospective validation–iteration, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus mengirim panel terpilih untuk assay konfirmasi dan memperbarui model dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas merancang siklus design-make-test-analyze sederhana sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama,

mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 8.9 validasi prospektif dan integrasi eksperimen sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang prospective validation akan memengaruhi cara hit rate dihitung, cara novelty dibandingkan, dan cara confirmation dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mengirim panel terpilih untuk assay konfirmasi dan memperbarui model', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi prospective validation, hit rate, novelty, confirmation, iteration dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
prospective validation	Bagaimana prospective validation didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
iteration	Bagaimana iteration memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$EF_x\% = (\text{hit pada x\% / jumlah pada x\%}) / (\text{total hit / total pustaka})$$

Enrichment factor

Praktikum 8 — Pipeline Penapisan Virtual Sederhana

Menggabungkan filter sifat, kemiripan ligan, peringkat konsensus, dan profil ADMET; docking dijelaskan sebagai opsi lanjutan.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
def property_filter(row):
    return (row["MW"] <= 500 and row["logP"] <= 5 and
            row["HBD"] <= 5 and row["HBA"] <= 10)
```

```
filtered = descriptor_table[descriptor_table.apply(property_filter,
axis=1)]
ranked = filtered.sort_values(["similarity", "predicted_activity"],
ascending=False)
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Penapisan Virtual dan ADMET harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara HTS dan artifact pada konteks kontrol positif dan negatif untuk menilai kualitas plate.
2. Identifikasi minimal tiga risiko jika peneliti menggunakan hit tunggal tanpa konfirmasi orthogonal.
3. Jelaskan dengan contoh hubungan antara heatmap dan SAR pada konteks klaster hit pada satu scaffold yang mungkin menunjukkan seri kimia.

4. Identifikasi minimal tiga risiko jika peneliti menyimpulkan aktivitas spesifik dari pola yang disebabkan interferensi.
5. Jelaskan dengan contoh hubungan antara Lipinski dan triage pada konteks menandai senyawa reaktif untuk tinjauan manual.
6. Identifikasi minimal tiga risiko jika peneliti membuang semua pelanggar aturan walau tujuan bukan obat oral.
7. Jelaskan dengan contoh hubungan antara similarity search dan ensemble pada konteks mencari analog dan scaffold hopping dari ligan aktif.
8. Identifikasi minimal tiga risiko jika peneliti menggunakan satu metode dan satu referensi tanpa menguji bias.
9. Jelaskan dengan contoh hubungan antara protein preparation dan cofactor pada konteks preparasi protease dengan residu katalitik terprotonasi tepat.
10. Identifikasi minimal tiga risiko jika peneliti mendocking pada struktur tanpa memeriksa ligan ko-kristal dan residu hilang.
11. Jelaskan dengan contoh hubungan antara pose dan docking pada konteks redocking ligan ko-kristal untuk menguji protokol.
12. Identifikasi minimal tiga risiko jika peneliti menyamakan skor docking dari perangkat berbeda.
13. Jelaskan dengan contoh hubungan antara scoring function dan decoy pada konteks menggabungkan skor docking dengan kemiripan interaksi.
14. Identifikasi minimal tiga risiko jika peneliti memilih kandidat hanya dari skor paling negatif.
15. Jelaskan dengan contoh hubungan antara ADMET dan toxicity pada konteks prediksi BBB, hERG, dan kelarutan untuk prioritas awal.
16. Identifikasi minimal tiga risiko jika peneliti menggantikan uji eksperimen dengan prediksi in silico tanpa validasi.
17. Jelaskan dengan contoh hubungan antara prospective validation dan iteration pada konteks mengirim panel terpilih untuk assay konfirmasi dan memperbarui model.
18. Identifikasi minimal tiga risiko jika peneliti melaporkan hanya hit berhasil tanpa seluruh kandidat yang diuji.
19. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
20. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
21. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
22. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 9. KIMIA KOMBINATORIAL, DESAIN PUSTAKA, INFORMATIKA REAKSI, DAN OPTIMASI

Orientasi Bab

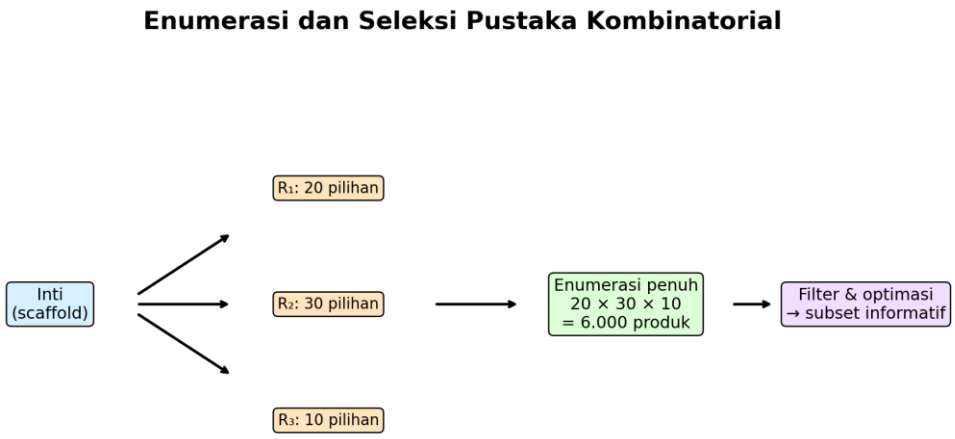
Bab ini menyusun fondasi konseptual dan praktis mengenai desain pustaka dan reaksi. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti desain pustaka dan reaksi dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

atom mapping, automation, building block, closed loop, condition prediction, constraint, coverage, disconnection, diverse library, DMTA, ELN, enrichment, enumeration, focused library, generative model, graph generation, LIMS, monomer-based, multiobjective, novelty, objective hacking, Pareto, product prediction, product property, product-based, R-group, reaction fingerprint, reaction rule, reagent, retrosynthesis, route search, sampling, scaffold, SELFIES, synthetic accessibility, target class, trade-off, utility, yield.



Gambar 9.1. Pertumbuhan kombinatorial dan kebutuhan filter serta seleksi informatif.

9.1 Prinsip Kimia Kombinatorial dan Enumerasi

Kombinasi scaffold, titik diversifikasi, dan building block dapat menghasilkan jutaan produk virtual sehingga aturan reaksi dan validasi struktur sangat penting. Pokok bahasan ini menempatkan scaffold, R-group, building block, enumeration, dan reaction rule sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah scaffold, R-group, building block, enumeration, dan reaction rule tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menghasilkan produk yang melanggar valensi atau tidak realistis secara sintesis. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah mengimplementasikan enumerasi kecil dengan reaction SMARTS. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap scaffold, R-group, building block, enumeration, dan reaction rule harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku:

informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
scaffold	Peran scaffold pada konteks Prinsip Kimia Kombinatorial dan Enumerasi.
R-group	Peran R-group pada konteks Prinsip Kimia Kombinatorial dan Enumerasi.
building block	Peran building block pada konteks Prinsip Kimia Kombinatorial dan Enumerasi.
enumeration	Peran enumeration pada konteks Prinsip Kimia Kombinatorial dan Enumerasi.
reaction rule	Peran reaction rule pada konteks Prinsip Kimia Kombinatorial dan Enumerasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 9.1 prinsip kimia kombinatorial dan enumerasi perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti scaffold dan hubungan operasionalnya dengan R-group. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu kombinasi scaffold, titik diversifikasi, dan building block dapat menghasilkan jutaan produk virtual sehingga aturan reaksi dan validasi struktur sangat penting, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan scaffold mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar building block. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang

dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah R-group, building block, dan enumeration harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus $20 \times 30 \times 10$ pilihan menghasilkan 6.000 produk sebelum filter, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan scaffold–reaction rule mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan scaffold, R-group, building block, enumeration, reaction rule sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 9.1 prinsip kimia kombinatorial dan enumerasi perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi scaffold, atau konversi R-group—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan building block sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menghasilkan produk yang melanggar valensi atau tidak realistis secara sintesis; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 9.1 prinsip kimia kombinatorial dan enumerasi dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan

yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.1 prinsip kimia kombinatorial dan enumerasi, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai scaffold–reaction rule, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas mengimplementasikan enumerasi kecil dengan reaction SMARTS sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.1 prinsip kimia kombinatorial dan enumerasi sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang scaffold akan memengaruhi cara R-group dihitung, cara building block dibandingkan, dan cara enumeration dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi scaffold, R-group, building block, enumeration, reaction rule dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
scaffold	Bagaimana scaffold didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reaction rule	Bagaimana reaction rule memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.2 Pustaka Beragam dan Pustaka Terfokus

Pustaka beragam bertujuan mencakup ruang kimia, sedangkan pustaka terfokus memperkaya wilayah yang relevan dengan target atau hipotesis tertentu. Pokok bahasan ini menempatkan diverse library, focused library, coverage, target class, dan enrichment sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah diverse library, focused library, coverage, target class, dan enrichment tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah pustaka kinase yang mempertahankan motif hinge-binding. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menilai pustaka hanya dari ukuran total. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak

valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan coverage dan enrichment dua desain. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap diverse library, focused library, coverage, target class, dan enrichment harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus pustaka kinase yang mempertahankan motif hinge-binding?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 9.2 pustaka beragam dan pustaka terfokus perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti diverse library dan hubungan operasionalnya dengan focused library. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu pustaka beragam bertujuan mencakup ruang kimia, sedangkan pustaka terfokus memperkaya wilayah yang relevan dengan target atau hipotesis tertentu, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan diverse library mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar coverage. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah focused library, coverage, dan target class harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus pustaka kinase yang mempertahankan motif hinge-binding, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan diverse library–enrichment mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan diverse library, focused library, coverage, target class, enrichment sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 9.2 pustaka beragam dan pustaka terfokus perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk pustaka kinase yang mempertahankan motif hinge-binding, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi diverse library, atau konversi focused library—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada

kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan coverage sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menilai pustaka hanya dari ukuran total; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 9.2 pustaka beragam dan pustaka terfokus dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk pustaka kinase yang mempertahankan motif hinge-binding, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.2 pustaka beragam dan pustaka terfokus, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk pustaka kinase yang mempertahankan motif hinge-binding, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris

benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai diverse library–enrichment, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus pustaka kinase yang mempertahankan motif hinge-binding dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas membandingkan coverage dan enrichment dua desain sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.2 pustaka beragam dan pustaka terfokus sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang diverse library akan memengaruhi cara focused library dihitung, cara coverage dibandingkan, dan cara target class dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada pustaka kinase yang mempertahankan motif hinge-binding', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi diverse library, focused library, coverage, target class, enrichment dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus

entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
diverse library	Bagaimana diverse library didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
enrichment	Bagaimana enrichment memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.3 Seleksi Monomer dan Seleksi Berbasis Produk

Seleksi monomer lebih murah tetapi tidak selalu mewakili sifat produk; seleksi berbasis produk lebih langsung namun memerlukan enumerasi atau sampling. Pokok bahasan ini menempatkan monomer-based, product-based, reagent, product property, dan sampling sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Sebagai bagian dari alur kerja reprodusibel, istilah monomer-based, product-based, reagent, product property, dan sampling tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer.

Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah building block lipofilik dapat menghasilkan produk terlalu hidrofobik. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih reagen secara independen tanpa melihat kombinasi akhir. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak

menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan dua strategi pada pustaka kecil. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap monomer-based, product-based, reagent, product property, dan sampling harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
monomer-based	Peran monomer-based pada konteks Seleksi Monomer dan Seleksi Berbasis Produk.
product-based	Peran product-based pada konteks Seleksi Monomer dan Seleksi Berbasis Produk.
reagent	Peran reagent pada konteks Seleksi Monomer dan Seleksi Berbasis Produk.
product property	Peran product property pada konteks Seleksi Monomer dan Seleksi Berbasis Produk.
sampling	Peran sampling pada konteks Seleksi Monomer dan Seleksi Berbasis Produk.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus building block lipofilik dapat menghasilkan produk terlalu hidrofobik?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 9.3 seleksi monomer dan seleksi berbasis produk perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti monomer-based dan hubungan operasionalnya dengan product-based. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang

sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu seleksi monomer lebih murah tetapi tidak selalu mewakili sifat produk; seleksi berbasis produk lebih langsung namun memerlukan enumerasi atau sampling, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan monomer-based mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar reagent. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah product-based, reagent, dan product property harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus building block lipofilik dapat menghasilkan produk terlalu hidrofobik, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan monomer-based–sampling mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan monomer-based, product-based, reagent, product property, sampling sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya,

pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 9.3 seleksi monomer dan seleksi berbasis produk perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk building block lipofilik dapat menghasilkan produk terlalu hidrofobik, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi monomer-based, atau konversi product-based—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan reagent sebelum koreksi struktur dapat mempertahankan

kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih reagen secara independen tanpa melihat kombinasi akhir; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 9.3 seleksi monomer dan seleksi berbasis produk dapat dipahami melalui empat komponen: masukan, transformasi,

parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk building block lipofilik dapat menghasilkan produk terlalu hidrofobik, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.3 seleksi monomer dan seleksi berbasis produk, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk building block lipofilik dapat menghasilkan produk terlalu hidrofobik, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai monomer-based–sampling, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus building block lipofilik dapat menghasilkan produk terlalu hidrofobik dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas membandingkan dua strategi pada pustaka kecil sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.3 seleksi monomer dan seleksi berbasis produk sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang monomer-based akan memengaruhi cara product-based dihitung, cara reagent dibandingkan, dan cara product property dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada building block lipofilik dapat menghasilkan produk terlalu hidrofobik', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi monomer-based, product-based, reagent, product property, sampling dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas

kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
monomer-based	Bagaimana monomer-based didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
sampling	Bagaimana sampling memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.4 Desain Multiobjektif dan Front Pareto

Aktivitas, selektivitas, admet, novelty, biaya, keragaman, dan kemudahan sintesis sering saling bertentangan sehingga front pareto lebih informatif daripada satu skor. Pokok bahasan ini menempatkan multiobjective, Pareto, trade-off, constraint, dan utility sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah multiobjective, Pareto, trade-off, constraint, dan utility tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memberi bobot tunggal tanpa analisis sensitivitas. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

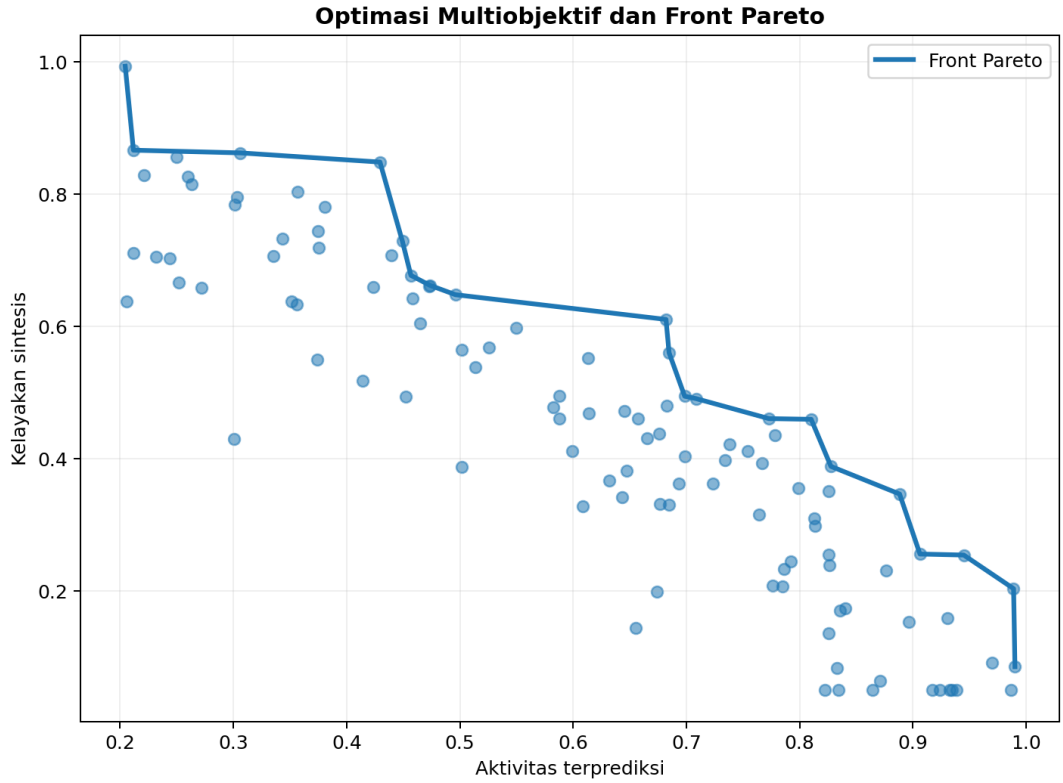
Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat plot Pareto dua tujuan. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap multiobjective, Pareto, trade-off, constraint, dan utility harus selalu dihubungkan dengan bab lain. Representasi

memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 9.2. Front Pareto menunjukkan solusi nondominan pada dua tujuan yang bertentangan.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 9.4 desain multiobjektif dan front pareto perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti multiobjective dan hubungan operasionalnya dengan Pareto. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas,

identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu aktivitas, selektivitas, ADMET, novelty, biaya, keragaman, dan kemudahan sintesis sering saling bertentangan sehingga front Pareto lebih informatif daripada satu skor, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan multiobjective mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar trade-off. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah Pareto, trade-off, dan constraint harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan multiobjective–utility mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan multiobjective, Pareto, trade-off, constraint, utility sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman

mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 9.4 desain multiobjektif dan front pareto perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi multiobjective, atau konversi Pareto—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran;

perhitungan trade-off sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memberi bobot tunggal tanpa analisis sensitivitas; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 9.4 desain multiobjektif dan front pareto dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan,

stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.4 desain multiobjektif dan front pareto, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model,

dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai multiobjective–utility, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas membuat plot Pareto dua tujuan sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.4 desain multiobjektif dan front pareto sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang multiobjective akan memengaruhi cara Pareto dihitung, cara trade-off dibandingkan, dan cara constraint dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi multiobjective, Pareto, trade-off, constraint, utility dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun

skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
multiobjective	Bagaimana multiobjective didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
utility	Bagaimana utility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.5 Informatika Reaksi dan Prediksi Hasil

Representasi reaksi, atom mapping, klasifikasi transformasi, prediksi produk, kondisi, dan rendemen memperluas informatika kimia ke perencanaan sintesis. Pokok bahasan ini menempatkan reaction fingerprint, atom mapping, product prediction, condition prediction, dan yield sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah reaction fingerprint, atom mapping, product prediction, condition prediction, dan yield tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data

dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah memprediksi produk utama dari reaktan dan kondisi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengabaikan data negatif dan ketidaklengkapan kondisi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun fitur reaksi untuk model rendemen. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap reaction fingerprint, atom mapping, product prediction, condition prediction, dan yield harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
reaction fingerprint	Peran reaction fingerprint pada konteks Informatika Reaksi dan Prediksi Hasil.
atom mapping	Peran atom mapping pada konteks Informatika Reaksi dan Prediksi Hasil.
product prediction	Peran product prediction pada konteks Informatika Reaksi dan Prediksi Hasil.
condition prediction	Peran condition prediction pada konteks Informatika Reaksi dan Prediksi Hasil.
yield	Peran yield pada konteks Informatika Reaksi dan Prediksi Hasil.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus memprediksi produk utama dari reaktan dan kondisi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif komputasi, unit analisis pada 9.5 informatika reaksi dan prediksi hasil perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti reaction fingerprint dan hubungan operasionalnya dengan atom mapping. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pada tahap verifikasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu representasi reaksi, atom mapping, klasifikasi transformasi, prediksi produk, kondisi, dan rendemen memperluas informatika kimia ke perencanaan sintesis, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan reaction fingerprint mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar product prediction. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Untuk pembelajaran berbasis proyek, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah atom mapping, product prediction, dan condition prediction harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus memprediksi produk utama dari reaktan dan kondisi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam skenario industri, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan reaction fingerprint–yield mengubah jumlah entitas unik serta interpretasi hasil.

Dalam riset akademik, peta konsep pada bagian ini dapat dibangun dengan menempatkan reaction fingerprint, atom mapping, product prediction, condition prediction, yield sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pertama, sumber data untuk 9.5 informatika reaksi dan prediksi hasil perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk memprediksi produk utama dari reaktan dan kondisi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Selanjutnya, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi reaction fingerprint, atau konversi atom mapping—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada sisi lain, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan product prediction sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus

cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam analisis yang lebih rinci, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengabaikan data negatif dan ketidaklengkapan kondisi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif kimia, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif komputasi, algoritme pada 9.5 informatika reaksi dan prediksi hasil dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus

mempunyai arti kimia yang dapat diinterpretasikan. Untuk memprediksi produk utama dari reaktan dan kondisi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pada tahap verifikasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Untuk pembelajaran berbasis proyek, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam skenario industri, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam riset akademik, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pertama, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.5 informatika reaksi dan prediksi hasil, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Selanjutnya, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk memprediksi produk utama dari reaktan dan kondisi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada sisi lain, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam analisis yang lebih rinci, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai reaction fingerprint–yield, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif kimia, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif komputasi, studi kasus memprediksi produk utama dari reaktan dan kondisi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pada tahap verifikasi, aktivitas menyusun fitur reaksi untuk model rendemen sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Untuk pembelajaran berbasis proyek, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam skenario industri, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam riset akademik, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.5 informatika reaksi dan prediksi hasil sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang reaction fingerprint akan memengaruhi cara atom mapping dihitung, cara product prediction dibandingkan, dan cara condition prediction dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada memprediksi produk utama dari reaktan dan kondisi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi reaction fingerprint, atom mapping, product prediction, condition prediction, yield dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas

kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
reaction fingerprint	Bagaimana reaction fingerprint didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
yield	Bagaimana yield memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.6 Retrosintesis Berbantuan Komputer

Retrosintesis memecah target menjadi prekursor melalui aturan atau model terpelajar, lalu mengevaluasi rute berdasarkan kelayakan, biaya, dan ketersediaan. Pokok bahasan ini menempatkan retrosynthesis, disconnection, route search, building block, dan synthetic accessibility sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah retrosynthesis, disconnection, route search, building block, dan synthetic accessibility tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika

hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mencari rute tiga tahap menuju analog sederhana. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap rute dengan skor tinggi otomatis dapat dikerjakan di laboratorium. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menilai rute dengan checklist kimia dan logistik. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap retrosynthesis, disconnection, route search, building block, dan synthetic accessibility harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mencari rute tiga tahap menuju analog sederhana?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pertama, unit analisis pada 9.6 retrosintesis berbantuan komputer perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti retrosynthesis dan hubungan operasionalnya dengan disconnection. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Selanjutnya, representasi selalu bersifat selektif. Fokus bagian ini, yaitu retrosintesis memecah target menjadi prekursor melalui aturan atau model terpelajar, lalu mengevaluasi rute berdasarkan kelayakan, biaya, dan ketersediaan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan retrosynthesis mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar route search. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada sisi lain, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah disconnection, route search, dan building block harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mencari rute tiga tahap menuju analog sederhana, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan

definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam analisis yang lebih rinci, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan retrosynthesis–synthetic accessibility mengubah jumlah entitas unik serta interpretasi hasil.

Dari perspektif kimia, peta konsep pada bagian ini dapat dibangun dengan menempatkan retrosynthesis, disconnection, route search, building block, synthetic accessibility sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif komputasi, sumber data untuk 9.6 retrosintesis berbantuan komputer perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak

otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mencari rute tiga tahap menuju analog sederhana, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada tahap verifikasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi retrosynthesis, atau konversi disconnection—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Untuk pembelajaran berbasis proyek, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan route search sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dalam skenario industri, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap rute dengan skor tinggi otomatis dapat dikerjakan di laboratorium; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam riset akademik, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pertama, algoritme pada 9.6 retrosintesis berbantuan komputer dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mencari rute tiga tahap menuju analog sederhana, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Selanjutnya, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada sisi lain, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline

cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam analisis yang lebih rinci, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dari perspektif kimia, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif komputasi, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.6 retrosintesis

berbantuan komputer, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada tahap verifikasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mencari rute tiga tahap menuju analog sederhana, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Untuk pembelajaran berbasis proyek, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dalam skenario industri, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai retrosynthesis–synthetic accessibility, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam riset akademik, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pertama, studi kasus mencari rute tiga tahap menuju analog sederhana dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Selanjutnya, aktivitas menilai rute dengan checklist kimia dan logistik sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada sisi lain, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam analisis yang lebih rinci, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dari perspektif kimia, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.6 retrosintesis berbantuan komputer sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang retrosynthesis akan memengaruhi cara disconnection dihitung, cara route search dibandingkan, dan cara building block dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mencari rute tiga tahap menuju analog sederhana', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan

mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi retrosynthesis, disconnection, route search, building block, synthetic accessibility dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
retrosynthesis	Bagaimana retrosynthesis didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
synthetic accessibility	Bagaimana synthetic accessibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.7 Model Generatif Molekul dan Kendala Kimia

Model berbasis smiles, selfies, graf, diffusion, atau reinforcement learning dapat mengusulkan molekul, tetapi validitas, novelty, synthesizability, dan objective hacking harus dikendalikan. Pokok bahasan ini menempatkan generative model, SELFIES, graph generation, novelty, dan objective hacking sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Untuk menjaga interpretasi kimia, istilah generative model, SELFIES, graph generation, novelty, dan objective hacking tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menghasilkan analog yang memenuhi rentang MW dan logP. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengoptimasi skor prediktif sampai menghasilkan struktur tidak realistis. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik,

pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat aturan validasi kandidat generatif. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap generative model, SELFIES, graph generation, novelty, dan objective hacking harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
generative model	Peran generative model pada konteks Model Generatif Molekul dan Kendala Kimia.
SELFIES	Peran SELFIES pada konteks Model Generatif Molekul dan Kendala Kimia.
graph generation	Peran graph generation pada konteks Model Generatif Molekul dan Kendala Kimia.
novelty	Peran novelty pada konteks Model Generatif Molekul dan Kendala Kimia.
objective hacking	Peran objective hacking pada konteks Model Generatif Molekul dan Kendala Kimia.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menghasilkan analog yang memenuhi rentang MW dan logP?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 9.7 model generatif molekul dan kendala kimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti generative model dan hubungan operasionalnya dengan SELFIES. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model berbasis SMILES, SELFIES, graf, diffusion, atau reinforcement learning dapat mengusulkan molekul, tetapi validitas, novelty, synthesizability, dan objective hacking harus dikendalikan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan generative model mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar graph generation. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah SELFIES, graph generation, dan novelty harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menghasilkan analog yang memenuhi rentang MW dan logP, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua

struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan generative model–objective hacking mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan generative model, SELFIES, graph generation, novelty, objective hacking sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 9.7 model generatif molekul dan kendala kimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menghasilkan analog yang memenuhi rentang MW dan logP, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi generative model, atau konversi SELFIES—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan graph generation sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengoptimasi skor prediktif sampai menghasilkan struktur tidak realistis; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 9.7 model generatif molekul dan kendala kimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menghasilkan analog yang memenuhi rentang MW dan logP, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi

yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.7 model generatif molekul dan kendala kimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menghasilkan analog yang memenuhi rentang MW dan logP, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai generative model–objective hacking, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.

Komponen audit	Pertanyaan kunci	Bukti minimum
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus menghasilkan analog yang memenuhi rentang MW dan logP dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas membuat aturan validasi kandidat generatif sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.7 model generatif molekul dan kendala kimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang generative model akan memengaruhi cara SELFIES dihitung, cara graph generation dibandingkan, dan cara novelty dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menghasilkan analog yang memenuhi rentang MW dan logP', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi,

sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi generative model, SELFIES, graph generation, novelty, objective hacking dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
generative model	Bagaimana generative model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
objective hacking	Bagaimana objective hacking memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

9.8 Otomasi Laboratorium, ELN, LIMS, dan Design-Make-Test-Analyze

Integrasi desain komputasi dengan inventori, robotika, catatan elektronik, dan hasil assay mempercepat iterasi jika standar data dan tanggung jawab jelas. Pokok

bahasan ini menempatkan ELN, LIMS, automation, DMTA, dan closed loop sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah ELN, LIMS, automation, DMTA, dan closed loop tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah otomasi proses yang belum distandardisasi sehingga kesalahan tersebar lebih cepat. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh,

menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menggambar alur data laboratorium digital. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap ELN, LIMS, automation, DMTA, dan closed loop harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 9.8 otomasi laboratorium, eln, lims, dan design-make-test-analyze perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti ELN dan hubungan operasionalnya dengan LIMS. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas,

identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu integrasi desain komputasi dengan inventori, robotika, catatan elektronik, dan hasil assay mempercepat iterasi jika standar data dan tanggung jawab jelas, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan ELN mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar automation. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah LIMS, automation, dan DMTA harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan ELN–closed loop mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan ELN, LIMS, automation, DMTA, closed loop sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 9.8 otomasi laboratorium, eln, lims, dan design-make-test-analyze perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi ELN, atau konversi LIMS—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan automation sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi.

Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa otomasi proses yang belum distandardisasi sehingga kesalahan tersebar lebih cepat; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 9.8 otomasi laboratorium, eln, lims, dan design-make-test-analyze dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta

prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 9.8 otomasi laboratorium, eln, lims, dan design-make-test-analyze, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya

dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai ELN–closed loop, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima

memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas menggambar alur data laboratorium digital sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 9.8 otomasi laboratorium, eln, lims, dan design-make-test-analyze sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang ELN akan memengaruhi cara LIMS dihitung, cara automation dibandingkan, dan cara DMTA dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi ELN, LIMS, automation, DMTA, closed loop dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu

menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
ELN	Bagaimana ELN didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
closed loop	Bagaimana closed loop memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Persamaan dan Metrik Penting

$$N_{produk} = \prod_i n_i \text{ untuk } n_i \text{ pilihan pada tiap titik diversifikasi.}$$

Ukuran enumerasi

Praktikum 9 — Enumerasi Pustaka dan Optimasi Pareto

Menggunakan reaction SMARTS untuk membuat pustaka kecil, memfilter produk, dan memilih front Pareto aktivitas-kelayakan.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.

- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
from rdkit import Chem
from rdkit.Chem import rdChemReactions

rxn = rdChemReactions.ReactionFromSmarts(
    "[C:1](=[O:2])Cl.[N:3]>>[C:1](=[O:2])[N:3]"
)
acid_chlorides = [Chem.MolFromSmiles("CC(=O)Cl"),
Chem.MolFromSmiles("O=C(Cl)c1ccccc1")]
amines = [Chem.MolFromSmiles("NCC"), Chem.MolFromSmiles("Nc1ccccc1")]
products = []
for a in acid_chlorides:
    for b in amines:
        outcomes = rxn.RunReactants((a, b))
        if outcomes:
            products.append(Chem.MolToSmiles(outcomes[0][0]))
print(sorted(set(products)))
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.
- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Desain Pustaka dan Reaksi harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara scaffold dan reaction rule pada konteks 20 × 30 × 10 pilihan menghasilkan 6.000 produk sebelum filter.
2. Identifikasi minimal tiga risiko jika peneliti menghasilkan produk yang melanggar valensi atau tidak realistis secara sintesis.
3. Jelaskan dengan contoh hubungan antara diverse library dan enrichment pada konteks pustaka kinase yang mempertahankan motif hinge-binding.
4. Identifikasi minimal tiga risiko jika peneliti menilai pustaka hanya dari ukuran total.
5. Jelaskan dengan contoh hubungan antara monomer-based dan sampling pada konteks building block lipofilik dapat menghasilkan produk terlalu hidrofobik.
6. Identifikasi minimal tiga risiko jika peneliti memilih reagen secara independen tanpa melihat kombinasi akhir.
7. Jelaskan dengan contoh hubungan antara multiobjective dan utility pada konteks kandidat sangat aktif tetapi sulit disintesis dibanding kandidat sedikit kurang aktif namun praktis.
8. Identifikasi minimal tiga risiko jika peneliti memberi bobot tunggal tanpa analisis sensitivitas.
9. Jelaskan dengan contoh hubungan antara reaction fingerprint dan yield pada konteks memprediksi produk utama dari reaktan dan kondisi.
10. Identifikasi minimal tiga risiko jika peneliti mengabaikan data negatif dan ketidaklengkapan kondisi.
11. Jelaskan dengan contoh hubungan antara retrosynthesis dan synthetic accessibility pada konteks mencari rute tiga tahap menuju analog sederhana.
12. Identifikasi minimal tiga risiko jika peneliti menganggap rute dengan skor tinggi otomatis dapat dikerjakan di laboratorium.

13. Jelaskan dengan contoh hubungan antara generative model dan objective hacking pada konteks menghasilkan analog yang memenuhi rentang MW dan logP.
14. Identifikasi minimal tiga risiko jika peneliti mengoptimasi skor prediktif sampai menghasilkan struktur tidak realistis.
15. Jelaskan dengan contoh hubungan antara ELN dan closed loop pada konteks pipeline memilih senyawa, memesan building block, mencatat sintesis, dan mengembalikan data.
16. Identifikasi minimal tiga risiko jika peneliti otomasi proses yang belum distandardisasi sehingga kesalahan tersebar lebih cepat.
17. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
18. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
19. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
20. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

BAB 10. PERANGKAT LUNAK, PEMROGRAMAN, REPRODUSIBILITAS, DAN ARAH MASA DEPAN

Orientasi Bab

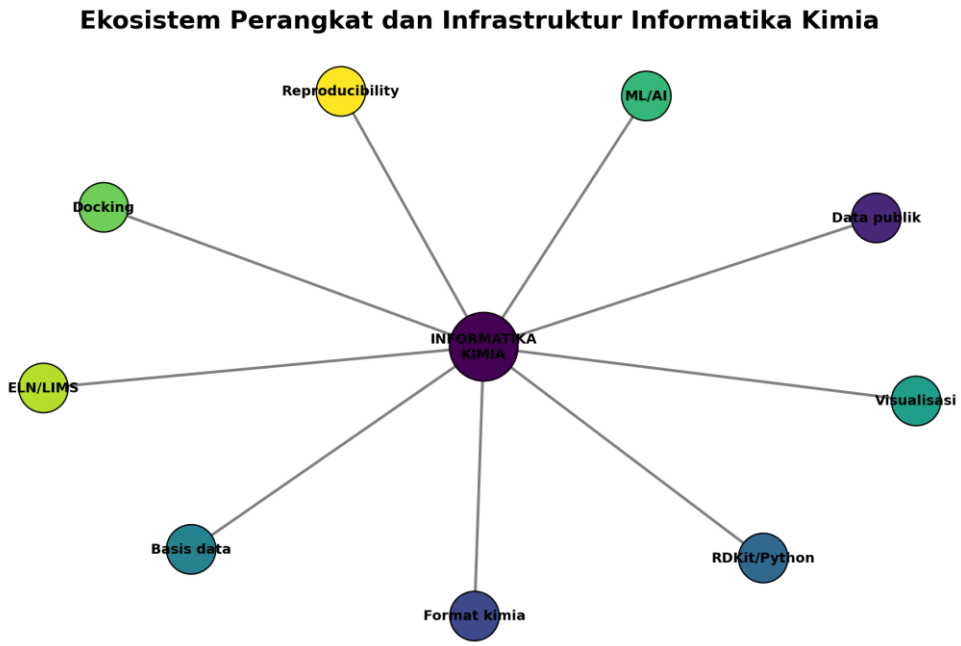
Bab ini menyusun fondasi konseptual dan praktis mengenai perangkat lunak dan masa depan. Struktur pembahasan mengikuti sintesis dua buku sumber utama, lalu diperluas dengan standar data, perangkat terbuka, pembelajaran mesin, dan praktik reprodusibilitas yang relevan untuk mahasiswa kimia S1.

Capaian Pembelajaran Bab

- Menjelaskan konsep inti perangkat lunak dan masa depan dengan istilah kimia dan komputasi yang tepat.
- Membedakan representasi, data, algoritme, asumsi, dan keluaran pada alur kerja terkait.
- Menerapkan prosedur dasar menggunakan data kecil yang dapat diperiksa secara manual.
- Mengevaluasi kualitas hasil, sumber bias, data leakage, dan batas penggunaan metode.
- Menyusun dokumentasi, visualisasi, dan laporan yang dapat direproduksi oleh kelompok lain.

Istilah Kunci

audit trail, cache, career, checksum, chemical language model, cloud, commercial, communication, container, DAG, data versioning, database cartridge, dependency, DOI, environment, foundation model, Git, human oversight, integration test, interdisciplinary, Jupyter, license, lifelong learning, logging, multimodal, open science, open source, pandas, pipeline, portfolio, Python, RDKit, repository, reproducibility, retrieval, scikit-learn, support, testing, toolkit, unit test, validation, vendor lock-in, visualization, workflow.



Gambar 10.1. Ekosistem perangkat, data, dan layanan yang dapat dirangkai sesuai kebutuhan.

10.1 Kategori Perangkat Informatika Kimia

Ekosistem perangkat mencakup editor struktur, toolkit, basis data, cartridge, visualisasi, workflow, qsar, docking, admet, eln, dan layanan awan. Pokok bahasan ini menempatkan toolkit, database cartridge, workflow, visualization, dan cloud sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah toolkit, database cartridge, workflow, visualization, dan cloud tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika

hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah memilih paket tunggal untuk seluruh masalah tanpa evaluasi interoperabilitas. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat matriks kebutuhan dan kategori alat. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap toolkit, database cartridge, workflow, visualization, dan cloud harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
toolkit	Peran toolkit pada konteks Kategori Perangkat Informatika Kimia.
database cartridge	Peran database cartridge pada konteks Kategori Perangkat Informatika Kimia.
workflow	Peran workflow pada konteks Kategori Perangkat Informatika Kimia.
visualization	Peran visualization pada konteks Kategori Perangkat Informatika Kimia.
cloud	Peran cloud pada konteks Kategori Perangkat Informatika Kimia.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 10.1 kategori perangkat informatika kimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti toolkit dan hubungan operasionalnya dengan database cartridge. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu ekosistem perangkat mencakup editor struktur, toolkit, basis data, cartridge, visualisasi, workflow, QSAR, docking, ADMET, ELN, dan layanan awan, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan toolkit mungkin sangat efisien untuk

pencarian, tetapi belum tentu memadai untuk menalar workflow. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah database cartridge, workflow, dan visualization harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan toolkit–cloud mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan toolkit, database cartridge, workflow, visualization, cloud sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 10.1 kategori perangkat informatika kimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi toolkit, atau konversi database cartridge—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan workflow sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa

apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa memilih paket tunggal untuk seluruh masalah tanpa evaluasi interoperabilitas; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 10.1 kategori perangkat informatika kimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.1 kategori perangkat informatika kimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster.

Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai toolkit–cloud, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas membuat matriks kebutuhan dan kategori alat sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data

mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.1 kategori perangkat informatika kimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang toolkit akan memengaruhi cara database cartridge dihitung, cara workflow dibandingkan, dan cara visualization dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi toolkit, database cartridge, workflow, visualization, cloud dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
toolkit	Bagaimana toolkit didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
cloud	Bagaimana cloud memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.2 Open Source, Komersial, dan Strategi Hibrida

Open source memberi transparansi dan fleksibilitas, komersial menawarkan dukungan dan integrasi, sedangkan strategi hibrida sering paling realistis. Pokok bahasan ini menempatkan open source, commercial, license, support, dan vendor lock-in sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dalam praktik penelitian, istilah open source, commercial, license, support, dan vendor lock-in tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah prototipe dengan open source lalu deployment dengan infrastruktur terkelola. Contoh perlu

dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengabaikan biaya total, lisensi, dan keberlanjutan. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membandingkan opsi berdasarkan kriteria institusi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap open source, commercial, license, support, dan vendor lock-in harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus prototipe dengan open source lalu deployment dengan infrastruktur terkelola?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam skenario industri, unit analisis pada 10.2 open source, komersial, dan strategi hibrida perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti open source dan hubungan operasionalnya dengan commercial. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam riset akademik, representasi selalu bersifat selektif. Fokus bagian ini, yaitu open source memberi transparansi dan fleksibilitas, komersial menawarkan dukungan dan integrasi, sedangkan strategi hibrida sering paling realistis, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan open source mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar license. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pertama, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah commercial, license, dan support harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus prototipe dengan open source lalu deployment dengan infrastruktur terkelola, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Selanjutnya, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua

struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan open source–vendor lock-in mengubah jumlah entitas unik serta interpretasi hasil.

Pada sisi lain, peta konsep pada bagian ini dapat dibangun dengan menempatkan open source, commercial, license, support, vendor lock-in sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam analisis yang lebih rinci, sumber data untuk 10.2 open source, komersial, dan strategi hibrida perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk prototipe dengan open source lalu deployment dengan infrastruktur terkelola, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut

memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif kimia, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi open source, atau konversi commercial—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif komputasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan license sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada tahap verifikasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengabaikan biaya total, lisensi, dan keberlanjutan; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Untuk pembelajaran berbasis proyek, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam skenario industri, algoritme pada 10.2 open source, komersial, dan strategi hibrida dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk prototipe dengan open source lalu deployment dengan infrastruktur terkelola, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam riset akademik, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pertama, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Selanjutnya, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada sisi lain, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam analisis yang lebih rinci, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.2 open source, komersial, dan strategi hibrida, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif kimia, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk prototipe dengan open source lalu deployment dengan infrastruktur terkelola, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif komputasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada tahap verifikasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai open source–vendor lock-in, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Untuk pembelajaran berbasis proyek, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.

Komponen audit	Pertanyaan kunci	Bukti minimum
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam skenario industri, studi kasus prototipe dengan open source lalu deployment dengan infrastruktur terkelola dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam riset akademik, aktivitas membandingkan opsi berdasarkan kriteria institusi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pertama, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Selanjutnya, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok

perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada sisi lain, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.2 open source, komersial, dan strategi hibrida sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang open source akan memengaruhi cara commercial dihitung, cara license dibandingkan, dan cara support dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada prototipe dengan open source lalu deployment dengan infrastruktur terkelola', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka

memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi open source, commercial, license, support, vendor lock-in dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open source	Bagaimana open source didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
vendor lock-in	Bagaimana vendor lock-in memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.3 Python, Notebook, dan RDKit sebagai Lingkungan Belajar

Python memudahkan integrasi data, komputasi, visualisasi, dan machine learning; notebook efektif untuk eksplorasi tetapi harus diubah menjadi pipeline yang dapat diuji. Pokok bahasan ini menempatkan Python, Jupyter, RDKit, pandas, dan scikit-learn sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya,

melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah Python, Jupyter, RDKit, pandas, dan scikit-learn tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menghitung deskriptor dan fingerprint dari SMILES. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah notebook dengan sel dijalankan tidak berurutan dan dependensi tidak tercatat. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat skrip dengan fungsi dan parameter yang jelas. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Python, Jupyter, RDKit, pandas, dan scikit-learn harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
Python	Peran Python pada konteks Python, Notebook, dan RDKit sebagai Lingkungan Belajar.
Jupyter	Peran Jupyter pada konteks Python, Notebook, dan RDKit sebagai Lingkungan Belajar.
RDKit	Peran RDKit pada konteks Python, Notebook, dan RDKit sebagai Lingkungan Belajar.
pandas	Peran pandas pada konteks Python, Notebook, dan RDKit sebagai Lingkungan Belajar.
scikit-learn	Peran scikit-learn pada konteks Python, Notebook, dan RDKit sebagai Lingkungan Belajar.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menghitung deskriptor dan fingerprint dari SMILES?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Untuk pembelajaran berbasis proyek, unit analisis pada 10.3 python, notebook, dan rdkit sebagai lingkungan belajar perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Python dan hubungan operasionalnya dengan Jupyter. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam skenario industri, representasi selalu bersifat selektif. Fokus bagian ini, yaitu Python memudahkan integrasi data, komputasi, visualisasi, dan machine learning; notebook efektif untuk eksplorasi tetapi harus diubah menjadi pipeline yang dapat diuji, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Python mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar RDKit. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam riset akademik, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah Jupyter, RDKit, dan pandas harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menghitung deskriptor dan fingerprint dari SMILES, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pertama, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Python–scikit-learn mengubah jumlah entitas unik serta interpretasi hasil.

Selanjutnya, peta konsep pada bagian ini dapat dibangun dengan menempatkan Python, Jupyter, RDKit, pandas, scikit-learn sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Pada sisi lain, sumber data untuk 10.3 python, notebook, dan rdkit sebagai lingkungan belajar perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menghitung deskriptor dan fingerprint dari SMILES, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam analisis yang lebih rinci, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Python, atau konversi Jupyter—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada

kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dari perspektif kimia, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan RDKit sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif komputasi, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa notebook dengan sel dijalankan tidak berurutan dan dependensi tidak tercatat; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Pada tahap verifikasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Untuk pembelajaran berbasis proyek, algoritme pada 10.3 python, notebook, dan rdkit sebagai lingkungan belajar dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk menghitung deskriptor dan fingerprint dari SMILES, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam skenario industri, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam riset akademik, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pertama, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode.

Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Selanjutnya, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Pada sisi lain, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.3 python, notebook, dan rdkit sebagai lingkungan belajar, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam analisis yang lebih rinci, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menghitung deskriptor dan

fingerprint dari SMILES, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dari perspektif kimia, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif komputasi, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Python–scikit-learn, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Pada tahap verifikasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Untuk pembelajaran berbasis proyek, studi kasus menghitung deskriptor dan fingerprint dari SMILES dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam skenario industri, aktivitas membuat skrip dengan fungsi dan parameter yang jelas sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam riset akademik, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pertama, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Selanjutnya, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk

mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.3 python, notebook, dan rdkit sebagai lingkungan belajar sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Python akan memengaruhi cara Jupyter dihitung, cara RDKit dibandingkan, dan cara pandas dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menghitung deskriptor dan fingerprint dari SMILES', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Python, Jupyter, RDKit, pandas, scikit-learn dengan data eksperimen, literatur, spektrum, atau target

biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Python	Bagaimana Python didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
scikit-learn	Bagaimana scikit-learn memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.4 Pipeline Data dan Workflow Otomatis

Workflow memisahkan tahap impor, kurasi, fitur, model, validasi, dan ekspor sehingga perubahan dapat ditelusuri dan diuji. Pokok bahasan ini menempatkan pipeline, DAG, cache, logging, dan testing sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah pipeline, DAG, cache, logging, dan testing tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah menjalankan analisis ulang otomatis ketika dataset berubah. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menyalin file manual antar-folder tanpa versioning. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

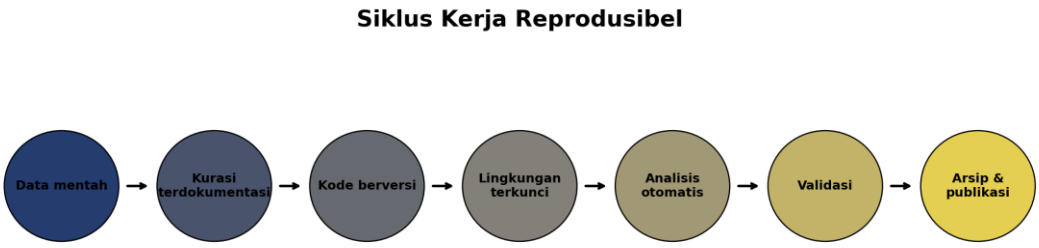
Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak

menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menggambar dependency graph pipeline. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap pipeline, DAG, cache, logging, dan testing harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.



Gambar 10.2. Siklus kerja reprodusibel dari data mentah sampai publikasi dan arsip.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus menjalankan analisis ulang otomatis ketika dataset berubah?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 10.4 pipeline data dan workflow otomatis perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti pipeline dan hubungan operasionalnya dengan DAG. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara.

Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu workflow memisahkan tahap impor, kurasi, fitur, model, validasi, dan ekspor sehingga perubahan dapat ditelusuri dan diuji, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan pipeline mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar cache. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah DAG, cache, dan logging harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus menjalankan analisis ulang otomatis ketika dataset berubah, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan pipeline–testing mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan pipeline, DAG, cache, logging, testing sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 10.4 pipeline data dan workflow otomatis perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk menjalankan analisis ulang otomatis ketika dataset berubah, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi pipeline, atau konversi DAG—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan cache sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci

sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menyalin file manual antar-folder tanpa versioning; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 10.4 pipeline data dan workflow otomatis dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus

mempunyai arti kimia yang dapat diinterpretasikan. Untuk menjalankan analisis ulang otomatis ketika dataset berubah, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.4 pipeline data dan workflow otomatis, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk menjalankan analisis ulang otomatis ketika dataset berubah, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai pipeline–testing, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus menjalankan analisis ulang otomatis ketika dataset berubah dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas menggambar dependency graph pipeline sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.4 pipeline data dan workflow otomatis sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang pipeline akan memengaruhi cara DAG dihitung, cara cache dibandingkan, dan cara logging dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada menjalankan analisis ulang otomatis ketika dataset berubah', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi pipeline, DAG, cache, logging, testing dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika

kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
pipeline	Bagaimana pipeline didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
testing	Bagaimana testing memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.5 Version Control, Environment, dan Container

Git, lock file, virtual environment, container, dan data versioning membantu memastikan orang lain dapat menjalankan analisis yang sama. Pokok bahasan ini menempatkan Git, environment, container, dependency, dan data versioning sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah Git, environment, container, dependency, dan data versioning tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data

awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah mencatat versi RDKit dan dataset dalam laporan. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menganggap kode saja cukup untuk reproduksi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah membuat README dengan langkah instalasi dan perintah eksekusi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap Git, environment, container, dependency, dan data versioning harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan

model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
Git	Peran Git pada konteks Version Control, Environment, dan Container.
environment	Peran environment pada konteks Version Control, Environment, dan Container.
container	Peran container pada konteks Version Control, Environment, dan Container.
dependency	Peran dependency pada konteks Version Control, Environment, dan Container.
data versioning	Peran data versioning pada konteks Version Control, Environment, dan Container.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus mencatat versi RDKit dan dataset dalam laporan?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 10.5 version control, environment, dan container perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti Git dan hubungan operasionalnya dengan environment. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu Git, lock file, virtual environment, container, dan data versioning membantu memastikan orang lain dapat menjalankan analisis yang sama, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan Git mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar container. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling

sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah environment, container, dan dependency harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus mencatat versi RDKit dan dataset dalam laporan, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan Git–data versioning mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan Git, environment, container, dependency, data versioning sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 10.5 version control, environment, dan container perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk mencatat versi RDKit dan dataset dalam laporan, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi Git, atau konversi environment—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan container sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menganggap kode saja cukup untuk reproduksi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 10.5 version control, environment, dan container dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk mencatat versi RDKit dan dataset dalam laporan, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan

yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.5 version control, environment, dan container, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk mencatat versi RDKit dan dataset dalam laporan, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai Git–data versioning, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus mencatat versi RDKit dan dataset dalam laporan dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas membuat README dengan langkah instalasi dan perintah eksekusi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap

klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.5 version control, environment, dan container sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang Git akan memengaruhi cara environment dihitung, cara container dibandingkan, dan cara dependency dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk

menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada mencatat versi RDKit dan dataset dalam laporan', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi Git, environment, container, dependency, data versioning dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
Git	Bagaimana Git didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
data versioning	Bagaimana data versioning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.6 Pengujian, Validasi Perangkat, dan Audit Trail

Unit test, integration test, checksum, validation set, dan audit log diperlukan terutama saat pipeline memengaruhi keputusan eksperimen. Pokok bahasan ini menempatkan unit test, integration test, checksum, audit trail, dan validation sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Pada tingkat implementasi, istilah unit test, integration test, checksum, audit trail, dan validation tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah uji bahwa standardisasi tidak mengubah jumlah atom berat. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengubah kode produksi tanpa pengujian regresi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun

tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menulis tiga test case untuk fungsi kurasi. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap unit test, integration test, checksum, audit trail, dan validation harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus uji bahwa standardisasi tidak mengubah jumlah atom berat?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada sisi lain, unit analisis pada 10.6 pengujian, validasi perangkat, dan audit trail perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti unit test dan hubungan operasionalnya dengan integration test. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dalam analisis yang lebih rinci, representasi selalu bersifat selektif. Fokus bagian ini, yaitu unit test, integration test, checksum, validation set, dan audit log diperlukan terutama saat pipeline memengaruhi keputusan eksperimen, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan unit test mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar checksum. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dari perspektif kimia, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah integration test, checksum, dan audit trail harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus uji bahwa standardisasi tidak mengubah jumlah atom berat, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dari perspektif komputasi, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan unit test–validation mengubah jumlah entitas unik serta interpretasi hasil.

Pada tahap verifikasi, peta konsep pada bagian ini dapat dibangun dengan menempatkan unit test, integration test, checksum, audit trail, validation sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Untuk pembelajaran berbasis proyek, sumber data untuk 10.6 pengujian, validasi perangkat, dan audit trail perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk uji bahwa standardisasi tidak mengubah jumlah atom berat, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dalam skenario industri, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi unit test, atau konversi integration test—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit

sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam riset akademik, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan checksum sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pertama, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengubah kode produksi tanpa pengujian regresi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Selanjutnya, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada sisi lain, algoritme pada 10.6 pengujian, validasi perangkat, dan audit trail dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk uji bahwa standardisasi tidak mengubah jumlah atom berat, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dalam analisis yang lebih rinci, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dari perspektif kimia, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dari perspektif komputasi, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari

metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pada tahap verifikasi, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Untuk pembelajaran berbasis proyek, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.6 pengujian, validasi perangkat, dan audit trail, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dalam skenario industri, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk uji bahwa standardisasi tidak

mengubah jumlah atom berat, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam riset akademik, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pertama, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai unit test–validation, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Selanjutnya, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada sisi lain, studi kasus uji bahwa standardisasi tidak mengubah jumlah atom berat dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dalam analisis yang lebih rinci, aktivitas menulis tiga test case untuk fungsi kurasi sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dari perspektif kimia, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dari perspektif komputasi, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pada tahap verifikasi, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk

mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.6 pengujian, validasi perangkat, dan audit trail sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang unit test akan memengaruhi cara integration test dihitung, cara checksum dibandingkan, dan cara audit trail dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada uji bahwa standardisasi tidak mengubah jumlah atom berat', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi unit test, integration test, checksum, audit trail, validation dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel,

tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
unit test	Bagaimana unit test didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
validation	Bagaimana validation memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.7 Kecerdasan Buatan Generatif dan Model Fondasi Kimia

Model bahasa kimia dan model multimodal dapat membantu pencarian, ekstraksi, desain, dan penjelasan, tetapi kualitas bergantung pada data, evaluasi, serta verifikasi manusia. Pokok bahasan ini menempatkan foundation model, multimodal, chemical language model, retrieval, dan human oversight sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah foundation model, multimodal, chemical language model, retrieval, dan human oversight tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah model membaca literatur dan menghubungkan struktur dengan kondisi reaksi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah menggunakan keluaran generatif tanpa pemeriksaan struktur dan referensi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model

modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah merancang protokol verifikasi hasil AI. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap foundation model, multimodal, chemical language model, retrieval, dan human oversight harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
foundation model	Peran foundation model pada konteks Kecerdasan Buatan Generatif dan Model Fondasi Kimia.
multimodal	Peran multimodal pada konteks Kecerdasan Buatan Generatif dan Model Fondasi Kimia.
chemical language model	Peran chemical language model pada konteks Kecerdasan Buatan Generatif dan Model Fondasi Kimia.
retrieval	Peran retrieval pada konteks Kecerdasan Buatan Generatif dan Model Fondasi Kimia.
human oversight	Peran human oversight pada konteks Kecerdasan Buatan Generatif dan Model Fondasi Kimia.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus model membaca literatur dan menghubungkan struktur dengan kondisi reaksi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dari perspektif kimia, unit analisis pada 10.7 kecerdasan buatan generatif dan model fondasi kimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat

berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti foundation model dan hubungan operasionalnya dengan multimodal. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Dari perspektif komputasi, representasi selalu bersifat selektif. Fokus bagian ini, yaitu model bahasa kimia dan model multimodal dapat membantu pencarian, ekstraksi, desain, dan penjelasan, tetapi kualitas bergantung pada data, evaluasi, serta verifikasi manusia, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan foundation model mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar chemical language model. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Pada tahap verifikasi, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah multimodal, chemical language model, dan retrieval harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus model membaca literatur dan menghubungkan struktur dengan kondisi reaksi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Untuk pembelajaran berbasis proyek, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan foundation model–human oversight mengubah jumlah entitas unik serta interpretasi hasil.

Dalam skenario industri, peta konsep pada bagian ini dapat dibangun dengan menempatkan foundation model, multimodal, chemical language model, retrieval, human oversight sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan

dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dalam riset akademik, sumber data untuk 10.7 kecerdasan buatan generatif dan model fondasi kimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk model membaca literatur dan menghubungkan struktur dengan kondisi reaksi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pertama, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi foundation model, atau konversi multimodal—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan

atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Selanjutnya, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan chemical language model sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Pada sisi lain, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa menggunakan keluaran generatif tanpa pemeriksaan struktur dan referensi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam analisis yang lebih rinci, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dari perspektif kimia, algoritme pada 10.7 kecerdasan buatan generatif dan model fondasi kimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk model membaca literatur dan menghubungkan struktur dengan kondisi reaksi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Dari perspektif komputasi, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Pada tahap verifikasi, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Untuk pembelajaran berbasis proyek, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam skenario industri, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang

relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dalam riset akademik, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.7 kecerdasan buatan generatif dan model fondasi kimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pertama, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk model membaca literatur dan menghubungkan struktur dengan kondisi reaksi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Selanjutnya, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran

domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Pada sisi lain, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai foundation model–human oversight, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam analisis yang lebih rinci, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dari perspektif kimia, studi kasus model membaca literatur dan menghubungkan struktur dengan kondisi reaksi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Dari perspektif komputasi, aktivitas merancang protokol verifikasi hasil AI sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Pada tahap verifikasi, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Untuk pembelajaran berbasis proyek, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam skenario industri, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.7 kecerdasan buatan generatif dan model fondasi kimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang foundation model akan memengaruhi cara multimodal dihitung, cara chemical language model dibandingkan, dan cara retrieval dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada model membaca literatur dan menghubungkan struktur dengan kondisi reaksi', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi foundation model, multimodal, chemical language model, retrieval, human oversight dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh

hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
foundation model	Bagaimana foundation model didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
human oversight	Bagaimana human oversight memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.8 Reproducible Research, Open Science, dan Publikasi Data

Publikasi ilmiah modern perlu menyertakan dataset, kode, environment, lisensi, dokumentasi, dan batasan yang memungkinkan audit serta penggunaan ulang. Pokok bahasan ini menempatkan open science, repository, DOI, license, dan reproducibility sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Ketika diterapkan pada koleksi molekul besar, istilah open science, repository, DOI, license, dan reproducibility tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan

kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah arsip dataset dan kode dengan DOI versi tetap. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah membagikan data tanpa metadata atau lisensi. Kekeliruan tersebut dapat menghasilkan analisis yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun paket reproduksi untuk satu praktikum. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap open science, repository, DOI, license, dan reproducibility harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus arsip dataset dan kode dengan DOI versi tetap?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Dalam riset akademik, unit analisis pada 10.8 reproducible research, open science, dan publikasi data perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti open science dan hubungan operasionalnya dengan repository. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Pertama, representasi selalu bersifat selektif. Fokus bagian ini, yaitu publikasi ilmiah modern perlu menyertakan dataset, kode, environment, lisensi, dokumentasi, dan batasan yang memungkinkan audit serta penggunaan ulang, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan open science mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar DOI. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Selanjutnya, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah repository, DOI, dan license harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus arsip dataset dan kode dengan DOI versi tetap, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Pada sisi lain, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan open science–reproducibility mengubah jumlah entitas unik serta interpretasi hasil.

Dalam analisis yang lebih rinci, peta konsep pada bagian ini dapat dibangun dengan menempatkan open science, repository, DOI, license, reproducibility sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Dari perspektif kimia, sumber data untuk 10.8 reproducible research, open science, dan publikasi data perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk arsip dataset dan kode dengan DOI versi tetap, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Dari perspektif komputasi, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi open science, atau konversi repository—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Pada tahap verifikasi, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan DOI sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Untuk pembelajaran berbasis proyek, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa membagikan data tanpa metadata atau lisensi; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dalam skenario industri, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier

yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Dalam riset akademik, algoritme pada 10.8 reproducible research, open science, dan publikasi data dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk arsip dataset dan kode dengan DOI versi tetap, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Pertama, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Selanjutnya, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Pada sisi lain, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Dalam analisis yang lebih rinci, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Dari perspektif kimia, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.8 reproducible research, open science, dan publikasi data, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Dari perspektif komputasi, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk arsip dataset dan kode dengan DOI versi tetap, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Pada tahap verifikasi, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Untuk pembelajaran berbasis proyek, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai open science–reproducibility, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dalam skenario industri, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi.

Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Dalam riset akademik, studi kasus arsip dataset dan kode dengan DOI versi tetap dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Pertama, aktivitas menyusun paket reproduksi untuk satu praktikum sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Selanjutnya, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama,

mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Pada sisi lain, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Dalam analisis yang lebih rinci, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.8 reproducible research, open science, dan publikasi data sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang open science akan memengaruhi cara repository dihitung, cara DOI dibandingkan, dan cara license dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini

efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada arsip dataset dan kode dengan DOI versi tetap', melainkan membandingkan dua kebijakan representasi atau dua strategi validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi open science, repository, DOI, license, reproducibility dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
open science	Bagaimana open science didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
reproducibility	Bagaimana reproducibility memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.

Komponen audit	Pertanyaan kunci	Bukti minimum
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

10.9 Peta Jalan Kompetensi dan Profesi Informatika Kimia

Karier bidang ini membutuhkan kombinasi kimia, pemrograman, statistika, basis data, komunikasi, dan kemampuan berkolaborasi lintas disiplin. Pokok bahasan ini menempatkan career, portfolio, interdisciplinary, communication, dan lifelong learning sebagai komponen yang harus dipahami secara terpadu. Dalam informatika kimia, suatu representasi atau algoritme tidak dinilai hanya dari kecanggihannya, melainkan dari kesesuaian antara pertanyaan kimia, bentuk data, asumsi matematis, dan keputusan yang hendak diambil. Karena itu, mahasiswa perlu selalu menanyakan apa yang dipertahankan oleh representasi, apa yang disederhanakan, dan informasi apa yang mungkin hilang.

Dari sudut pandang mahasiswa kimia, istilah career, portfolio, interdisciplinary, communication, dan lifelong learning tidak berdiri sendiri. Setiap istilah menentukan cara objek kimia dipetakan ke bentuk yang dapat diproses komputer. Perubahan kecil dalam aturan atom, ikatan, satuan, scaling, atau pembagian data dapat mengubah hasil analisis. Disiplin ini menuntut dua ketelitian sekaligus: ketelitian kimia untuk menjaga makna molekul dan ketelitian komputasi untuk menjaga konsistensi operasi. Keduanya harus terdokumentasi sehingga orang lain dapat menelusuri asal suatu hasil.

Alur kerja yang baik biasanya dimulai dengan definisi objek dan endpoint, dilanjutkan pemeriksaan identitas serta kualitas data, pemilihan representasi, penerapan metode, validasi, dan interpretasi. Pada setiap tahap, parameter penting dicatat dalam metadata. Dengan pola tersebut, kesalahan dapat dilokalisasi. Jika hasil berubah, peneliti dapat menentukan apakah penyebabnya terletak pada data awal, standardisasi, fitur, algoritme, atau prosedur evaluasi, bukan sekadar mengulangi perhitungan secara coba-coba.

Contoh yang dapat digunakan untuk membumikan konsep ini adalah portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi. Contoh perlu dibahas pada tiga tingkat. Tingkat pertama adalah struktur atau fenomena kimianya. Tingkat kedua adalah cara komputer menyimpan dan memprosesnya. Tingkat ketiga adalah konsekuensi ilmiah dari keluaran yang diperoleh. Pembahasan berlapis seperti ini mencegah mahasiswa terjebak pada penggunaan perangkat lunak tanpa memahami makna kimia dari input, proses, dan output.

Kesalahan yang sering muncul adalah mengukur kompetensi hanya dari jumlah perangkat yang pernah dipakai. Kekeliruan tersebut dapat menghasilkan analisis

yang tampak rapi namun tidak valid. Tanda peringatannya antara lain hasil yang terlalu baik, perubahan besar akibat parameter kecil, rekaman yang tidak dapat dilacak, dan kesimpulan yang melampaui domain data. Pencegahannya mencakup pemeriksaan kimia manual pada sampel, baseline sederhana, visualisasi diagnostik, pengujian sensitivitas, serta pembandingan dengan sumber atau metode independen.

Dalam pembelajaran S1, kedalaman materi sebaiknya bergerak dari pemahaman konsep menuju kemampuan melakukan analisis. Mahasiswa pertama-tama menjelaskan istilah dengan bahasanya sendiri, kemudian membaca contoh, menjalankan prosedur terbimbing, mengubah satu parameter, dan akhirnya mempertanggungjawabkan pilihan metode. Urutan ini lebih kuat daripada menghafal nama algoritme karena membentuk kebiasaan berpikir ilmiah dan keterampilan audit yang diperlukan pada proyek nyata.

Perkembangan mutakhir, termasuk basis data terbuka, komputasi awan, pembelajaran mesin, model graf, dan AI generatif, memperluas skala serta kecepatan analisis. Namun prinsip dasar pada bagian ini tetap berlaku. Model modern tetap membutuhkan identitas struktur yang benar, data yang terkurasi, split yang jujur, metrik yang tepat, serta pelaporan keterbatasan. Teknologi baru tidak menghapus kebutuhan validasi; sebaliknya, skala yang lebih besar membuat kesalahan kecil dapat menyebar lebih luas.

Aktivitas yang disarankan adalah menyusun rencana portofolio satu tahun. Hasil aktivitas tidak cukup berupa angka atau gambar. Mahasiswa perlu melampirkan data masukan, aturan yang dipakai, parameter, kode atau langkah kerja, tabel hasil, interpretasi kimia, dan catatan keterbatasan. Dengan demikian, artefak pembelajaran sekaligus menjadi portofolio yang menunjukkan penguasaan materi serta kesiapan bekerja dalam tim multidisiplin.

Sebagai simpulan bagian, pemahaman terhadap career, portfolio, interdisciplinary, communication, dan lifelong learning harus selalu dihubungkan dengan bab lain. Representasi memengaruhi deskriptor; deskriptor memengaruhi kemiripan dan model; kurasi memengaruhi seluruh keluaran; dan tujuan keputusan menentukan metode yang layak. Hubungan berantai ini merupakan tema utama buku: informatika kimia adalah sistem penalaran dan pengelolaan bukti, bukan sekadar kumpulan perangkat komputasi.

Konsep	Makna operasional dalam bagian ini
career	Peran career pada konteks Peta Jalan Kompetensi dan Profesi Informatika Kimia.
portfolio	Peran portfolio pada konteks Peta Jalan Kompetensi dan Profesi Informatika Kimia.
interdisciplinary	Peran interdisciplinary pada konteks Peta Jalan Kompetensi dan Profesi Informatika Kimia.

Konsep	Makna operasional dalam bagian ini
communication	Peran communication pada konteks Peta Jalan Kompetensi dan Profesi Informatika Kimia.
lifelong learning	Peran lifelong learning pada konteks Peta Jalan Kompetensi dan Profesi Informatika Kimia.

Pertanyaan Pemantik

Bagaimana perubahan pilihan representasi atau parameter pada bagian ini dapat mengubah keputusan untuk kasus portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi?

Pendalaman I — Objek, Makna, dan Tingkat Representasi

Pada tahap verifikasi, unit analisis pada 10.9 peta jalan kompetensi dan profesi informatika kimia perlu ditetapkan sebelum data dikumpulkan. Unit itu dapat berupa satu struktur, satu bentuk ionisasi, satu konformer, satu reaksi, satu hasil assay, atau satu pasangan molekul-target. Pilihan unit menentukan arti career dan hubungan operasionalnya dengan portfolio. Bila unit analisis berubah di tengah pekerjaan, tabel yang tampak konsisten dapat mencampurkan objek yang sebenarnya tidak setara. Karena itu, setiap proyek perlu mempunyai definisi entitas, identifier utama, aturan versi, dan kebijakan ketika satu senyawa mempunyai beberapa rekaman atau beberapa keadaan kimia.

Untuk pembelajaran berbasis proyek, representasi selalu bersifat selektif. Fokus bagian ini, yaitu karier bidang ini membutuhkan kombinasi kimia, pemrograman, statistika, basis data, komunikasi, dan kemampuan berkolaborasi lintas disiplin, mengharuskan peneliti menjelaskan atribut mana yang dipertahankan dan atribut mana yang diabaikan. Representasi yang menekankan career mungkin sangat efisien untuk pencarian, tetapi belum tentu memadai untuk menalar interdisciplinary. Sebaliknya, representasi yang kaya dapat memperbesar biaya penyimpanan, waktu komputasi, dan risiko inkonsistensi. Prinsip yang berguna adalah memakai representasi paling sederhana yang masih mempertahankan informasi yang dibutuhkan oleh keputusan, kemudian membandingkannya dengan representasi alternatif sebagai uji ketahanan.

Dalam skenario industri, makna kimia tidak boleh dikalahkan oleh kenyamanan format. Istilah portfolio, interdisciplinary, dan communication harus diterjemahkan ke aturan yang dapat diuji: bagaimana nilainya diperoleh, satuan apa yang digunakan, kondisi apa yang berlaku, dan kapan nilainya tidak dapat dibandingkan. Pada kasus portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi, satu nilai tanpa konteks dapat menyesatkan karena dua laboratorium atau dua perangkat mungkin menerapkan definisi, protokol, atau parameter yang

berbeda. Metadata bukan hiasan administratif; metadata adalah bagian dari observasi dan menentukan apakah data layak digabungkan, dimodelkan, atau digunakan kembali.

Dalam riset akademik, terdapat perbedaan antara identitas, kesetaraan, dan kemiripan. Dua rekaman dapat merujuk pada parent compound yang sama tetapi berbeda garam, stereoisomer, tautomer, isotop, atau keadaan protonasi. Pada saat yang sama, dua struktur berbeda dapat dianggap ekuivalen untuk tujuan tertentu, misalnya ketika analisis hanya memakai kerangka karbon. Hubungan tersebut perlu dinyatakan sebagai kebijakan, bukan dibiarkan menjadi perilaku bawaan perangkat. Mahasiswa sebaiknya menguji sedikitnya lima pasangan contoh dan mencatat bagaimana kebijakan career–lifelong learning mengubah jumlah entitas unik serta interpretasi hasil.

Pertama, peta konsep pada bagian ini dapat dibangun dengan menempatkan career, portfolio, interdisciplinary, communication, lifelong learning sebagai simpul, lalu menambahkan relasi 'mewakili', 'diturunkan dari', 'mempengaruhi', 'divalidasi oleh', dan 'membatasi'. Peta semacam itu membantu membedakan hubungan definisional dari hubungan empiris. Hubungan definisional mengikuti aturan representasi, sedangkan hubungan empiris memerlukan data dan pengujian. Ketika mahasiswa mampu menunjukkan jenis relasi dan bukti yang menopangnya, pemahaman mereka bergerak dari hafalan istilah menuju model mental yang dapat dipakai untuk mendiagnosis kasus baru.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman II — Akuisisi Data, Kurasi, dan Rekayasa Alur Kerja

Selanjutnya, sumber data untuk 10.9 peta jalan kompetensi dan profesi informatika kimia perlu dinilai pada empat lapis: otoritas sumber, cara pengukuran atau pembangkitan, kelengkapan metadata, dan kompatibilitas lisensi. Data publik tidak

otomatis bebas masalah. Rekaman dapat berasal dari agregasi sekunder, memiliki pembaruan yang berbeda, atau menggunakan identifier yang berubah. Untuk portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi, prosedur yang aman adalah menyimpan snapshot asli, tanggal akses, versi basis data, kueri yang digunakan, serta checksum berkas. Langkah tersebut memungkinkan peneliti membedakan perubahan akibat data sumber dari perubahan akibat kode analisis.

Pada sisi lain, kurasi sebaiknya dirancang sebagai serangkaian transformasi eksplisit, bukan penyuntingan manual tersembunyi. Setiap transformasi—misalnya normalisasi nama, penanganan nilai hilang, pemisahan campuran, standardisasi career, atau konversi portfolio—diberi nama, alasan, parameter, jumlah rekaman terdampak, dan keluaran antara. Tabel audit sebelum-sesudah sangat berguna karena menunjukkan apakah perubahan besar terjadi pada kelompok tertentu. Jika satu aturan menghapus sebagian besar senyawa bermuatan atau kelas struktur langka, konsekuensinya terhadap ruang kimia dan generalisasi perlu dibahas.

Dalam analisis yang lebih rinci, urutan operasi dapat mengubah hasil. Deduplicasi sebelum standardisasi mungkin menghasilkan jumlah entitas berbeda dibanding deduplikasi setelah standardisasi; imputasi sebelum split dapat menyebabkan kebocoran; perhitungan interdisciplinary sebelum koreksi struktur dapat mempertahankan kesalahan ke tahap berikutnya. Oleh sebab itu, pipeline perlu dipandang sebagai objek ilmiah yang mempunyai urutan, dependensi, dan versi. Diagram alir harus cukup rinci sehingga kelompok lain dapat mengimplementasikan kembali pipeline tanpa menebak langkah yang hilang.

Dari perspektif kimia, kontrol kualitas memerlukan kombinasi aturan otomatis dan inspeksi kimia. Aturan otomatis unggul untuk skala besar, sedangkan inspeksi manual diperlukan untuk memahami pola error dan memeriksa apakah aturan terlalu agresif. Sampel inspeksi sebaiknya mencakup rekaman acak, outlier, struktur kompleks, nilai ekstrem, serta kasus yang diubah atau dibuang oleh pipeline. Pada bagian ini, kesalahan khas berupa mengukur kompetensi hanya dari jumlah perangkat yang pernah dipakai; karena itu, laporan kurasi perlu memperlihatkan contoh konkret sebelum-sesudah dan bukan hanya jumlah akhir.

Dari perspektif komputasi, penyimpanan hasil kurasi hendaknya memisahkan data mentah, data antara, data final, dan laporan kualitas. Berkas final diberi skema kolom, tipe data, satuan, kamus nilai kategorik, kebijakan nilai hilang, dan identifier yang menghubungkannya ke sumber. Penggunaan format sederhana seperti CSV dapat diterima untuk tabel kecil, tetapi struktur, spektrum, atau relasi kompleks mungkin memerlukan SDF, JSON, basis data relasional, atau graf pengetahuan. Pilihan format harus mengikuti struktur data serta kebutuhan validasi, bukan sekadar kebiasaan perangkat lunak.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman III — Algoritme, Parameter, dan Kompleksitas Komputasi

Pada tahap verifikasi, algoritme pada 10.9 peta jalan kompetensi dan profesi informatika kimia dapat dipahami melalui empat komponen: masukan, transformasi, parameter, dan keluaran. Masukan perlu mempunyai tipe serta prasyarat yang jelas; transformasi menjelaskan operasi inti; parameter mengendalikan perilaku; dan keluaran harus mempunyai arti kimia yang dapat diinterpretasikan. Untuk portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi, mahasiswa perlu menuliskan pseudocode ringkas sebelum menjalankan perangkat. Pseudocode memaksa pemisahan antara langkah kimia, langkah numerik, serta langkah administratif seperti membaca dan menulis berkas.

Untuk pembelajaran berbasis proyek, parameter bukan sekadar tombol teknis. Threshold, radius, ukuran bit, jumlah konformer, fungsi jarak, nilai regularisasi, atau kriteria konvergensi mengandung asumsi mengenai variasi yang dianggap penting. Analisis sensitivitas dapat dilakukan dengan memilih rentang parameter yang masuk akal, menjalankan pipeline secara sistematis, dan membandingkan stabilitas peringkat, klaster, koefisien, atau keputusan. Bila perubahan kecil menghasilkan keputusan yang sangat berbeda, hasil perlu dilaporkan sebagai tidak stabil dan tidak boleh disederhanakan menjadi satu angka pasti.

Dalam skenario industri, kompleksitas komputasi menjadi relevan ketika koleksi bertambah. Operasi pasangan pada n molekul dapat tumbuh mendekati n kuadrat, pencarian subgraf dapat mahal, dan pembangkitan konformer atau enumerasi pustaka dapat meledak secara kombinatorial. Strategi yang tersedia meliputi screening awal, indeks, hashing, batching, paralelisasi, sampling, serta penggunaan representasi bertingkat. Akan tetapi, optimasi kecepatan harus diuji terhadap

akurasi. Pipeline cepat yang kehilangan kelas senyawa penting tidak lebih baik daripada pipeline lambat yang tidak dapat dijalankan.

Dalam riset akademik, implementasi perlu membedakan determinisme dan stokastisitas. Algoritme deterministik diharapkan memberi keluaran sama untuk input dan versi yang sama, sedangkan metode stokastik membutuhkan seed serta pengulangan. Bahkan metode deterministik dapat berubah ketika pustaka, model aromatisitas, urutan atom, perangkat keras, atau presisi numerik berubah. Karena itu, catatan versi, parameter, seed, dan lingkungan eksekusi merupakan bagian dari metode. Mahasiswa sebaiknya mengulang analisis pada lingkungan baru untuk melihat bagian mana yang benar-benar portabel.

Pertama, baseline membantu menilai apakah kerumitan tambahan mempunyai manfaat nyata. Baseline dapat berupa aturan kimia sederhana, statistik deskriptif, pencarian exact match, model linear, atau pemilihan acak yang dikontrol. Metode lebih kompleks baru dianggap bernilai bila meningkatkan ukuran yang relevan, stabil pada validasi, dan tidak mengorbankan interpretabilitas atau biaya secara tidak proporsional. Perbandingan harus menggunakan data, split, dan metrik yang sama. Dengan demikian, klaim kemajuan tidak bergantung pada perubahan kondisi evaluasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman IV — Validasi, Ketidakpastian, dan Diagnosis Kesalahan

Selanjutnya, validasi dimulai dengan menanyakan jenis klaim. Klaim identitas memerlukan verifikasi struktur; klaim prediksi memerlukan data uji yang independen; klaim cakupan memerlukan analisis ruang kimia; dan klaim kausal memerlukan desain yang jauh lebih kuat daripada korelasi. Pada 10.9 peta jalan

kompetensi dan profesi informatika kimia, metrik harus dipilih setelah klaim dan biaya kesalahan ditetapkan. Akurasi rata-rata tidak cukup bila false negative dan false positive mempunyai konsekuensi berbeda atau bila data sangat tidak seimbang.

Pada sisi lain, pemisahan data harus meniru penggunaan nyata. Split acak sering terlalu optimistis ketika analog dekat, seri sintesis, batch, atau sumber yang sama muncul di data latih dan uji. Alternatifnya adalah scaffold split, time split, group split, atau external test set. Untuk portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi, mahasiswa perlu menjelaskan unit independensi: apakah tiap baris benar-benar observasi independen, atau beberapa baris berasal dari molekul, eksperimen, dan laboratorium yang sama. Tanpa definisi ini, validasi dapat mengukur kemampuan mengingat alih-alih kemampuan menggeneralisasi.

Dalam analisis yang lebih rinci, ketidakpastian mempunyai beberapa sumber: variasi pengukuran, ketidaklengkapan sampling, pilihan representasi, parameter model, dan pergeseran domain. Satu interval numerik tidak selalu merangkum semuanya. Laporan yang baik memisahkan ketidakpastian aleatorik dari epistemik sejauh mungkin, menunjukkan variasi antar-split atau antar-seed, dan menandai prediksi di luar domain aplikasi. Untuk keputusan laboratorium, daftar kandidat sebaiknya dilengkapi alasan pemilihan, tingkat keyakinan, dan rencana eksperimen konfirmasi.

Dari perspektif kimia, diagnosis error memerlukan pemeriksaan pada tingkat agregat dan tingkat molekul. Pada tingkat agregat, gunakan distribusi residual, confusion matrix, kurva kalibrasi, learning curve, atau stabilitas klaster. Pada tingkat molekul, tampilkan struktur contoh benar dan salah, tetangga terdekat, nilai career–lifelong learning, serta kemungkinan penyebab kimia. Pola error dapat mengungkap label salah, mekanisme berbeda, keadaan ionisasi yang tidak sesuai, atau representasi yang gagal menangkap fitur penting.

Dari perspektif komputasi, ancaman validitas perlu ditulis secara eksplisit. Ancaman internal berkaitan dengan kurasi, kebocoran, parameter, dan pelaksanaan; ancaman eksternal berkaitan dengan generalisasi ke kelas molekul, kondisi, atau laboratorium lain; ancaman konstruk berkaitan dengan apakah endpoint benar-benar mengukur konsep yang dimaksud. Tindakan mitigasi tidak selalu menghilangkan ancaman, tetapi membuat pembaca mengetahui batas interpretasi. Transparansi ini lebih ilmiah daripada menyembunyikan kelemahan di balik nilai metrik yang tinggi.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman V — Studi Kasus, Keputusan, dan Portofolio Reprodusibel

Pada tahap verifikasi, studi kasus portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi dapat dirancang sebagai keputusan bertahap. Tahap pertama menetapkan pertanyaan dan kriteria sukses; tahap kedua mengumpulkan serta mengaudit data; tahap ketiga membandingkan representasi dan metode; tahap keempat melakukan validasi; dan tahap kelima memilih tindakan eksperimen. Setiap tahap menghasilkan artefak yang dapat diperiksa. Pendekatan bertahap mencegah keputusan akhir bergantung pada satu skor dan memungkinkan kelompok menghentikan proyek ketika kualitas data tidak memadai.

Untuk pembelajaran berbasis proyek, aktivitas menyusun rencana portofolio satu tahun sebaiknya menghasilkan lebih dari laporan naratif. Portofolio minimum terdiri atas berkas data mentah atau tautan permanen, data terkurasi, kamus data, notebook atau skrip, file environment, README, gambar, tabel, dan log keputusan. Notebook tidak boleh menjadi satu-satunya dokumentasi karena urutan eksekusi dan dependensi sering tidak jelas. Skrip utama atau pipeline perlu dapat dijalankan dari awal dengan satu perintah yang terdokumentasi.

Dalam skenario industri, komunikasi hasil harus menghubungkan angka dengan struktur kimia. Tabel menyajikan nilai dan ketidakpastian, grafik memperlihatkan pola, sedangkan gambar struktur memperlihatkan alasan kimia yang mungkin. Caption perlu menyatakan apa yang dibandingkan, data yang digunakan, dan pesan utama tanpa memaksa pembaca mencari penjelasan di paragraf lain. Untuk setiap klaim utama, mahasiswa dapat membuat matriks bukti yang menunjukkan data, analisis, visualisasi, serta batasan yang mendukung klaim tersebut.

Dalam riset akademik, kerja kelompok maksimal tiga mahasiswa dapat dibagi menjadi peran kurator data, pengembang analisis, dan penelaah kimia, tetapi seluruh anggota wajib memahami keseluruhan pipeline. Rotasi peran pada beberapa

iterasi mencegah ketergantungan pada satu orang. Log kontribusi mencatat keputusan, commit, pemeriksaan silang, dan masalah yang belum selesai. Penilaian kelompok perlu menggabungkan kualitas produk, kemampuan menjelaskan pilihan, serta kemampuan mereproduksi pekerjaan anggota lain.

Pertama, refleksi akhir perlu menjawab empat pertanyaan: informasi apa yang paling menentukan keputusan; asumsi apa yang paling berisiko; bukti apa yang dapat mengubah kesimpulan; dan eksperimen apa yang paling efisien untuk mengurangi ketidakpastian. Jawaban tersebut memperlihatkan bahwa informatika kimia tidak berhenti pada komputasi. Tujuannya adalah membangun rantai bukti yang dapat dipertanggungjawabkan, menghemat eksperimen yang tidak perlu, serta memilih eksperimen berikutnya secara lebih rasional.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Pendalaman VI — Integrasi Lintas Bab dan Agenda Penelitian

Integrasi lintas bab menempatkan 10.9 peta jalan kompetensi dan profesi informatika kimia sebagai mata rantai dalam sistem yang lebih besar. Keputusan tentang career akan memengaruhi cara portfolio dihitung, cara interdisciplinary dibandingkan, dan cara communication dimasukkan ke model atau basis data. Mahasiswa perlu melatih penalaran hulu-hilir: menelusuri dampak satu keputusan standardisasi hingga visualisasi dan kesimpulan, lalu menelusuri balik sebuah anomali hasil menuju sumber data serta representasi awal. Latihan penelusuran dua arah ini efektif untuk menemukan asumsi tersembunyi yang tidak tampak ketika tiap bab dipelajari secara terpisah.

Sebuah pertanyaan penelitian yang baik pada tema ini mengandung objek, konteks, pembanding, dan kriteria keberhasilan. Contohnya, bukan sekadar 'menerapkan metode pada portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi', melainkan membandingkan dua kebijakan representasi atau dua strategi

validasi dan mengukur pengaruhnya terhadap stabilitas keputusan. Pertanyaan semacam itu dapat dijawab oleh mahasiswa S1 karena ruang lingkupnya terukur, tetapi tetap memberi kontribusi metodologis. Data dan kode terbuka memungkinkan replikasi, sedangkan variasi dataset lokal dapat memberi konteks baru tanpa harus mengklaim penemuan yang melampaui bukti.

Agenda pengembangan berikutnya dapat diarahkan pada integrasi career, portfolio, interdisciplinary, communication, lifelong learning dengan data eksperimen, literatur, spektrum, atau target biologis. Tantangannya bukan hanya menggabungkan tabel, tetapi menyelaraskan identifier, ontologi, satuan, tingkat granularitas, dan ketidakpastian. Prototipe kecil lebih bernilai daripada sistem besar yang tidak dapat diaudit. Mahasiswa dapat memulai dengan sepuluh hingga seratus entitas, membangun skema yang jelas, menguji kueri atau analisis, lalu mengevaluasi apakah perluasan skala mempertahankan makna kimia serta kinerja komputasi.

Pada akhirnya, penguasaan bagian ini dibuktikan melalui kemampuan menjelaskan pilihan, bukan sekadar memperoleh keluaran. Mahasiswa harus mampu menunjukkan alasan pemilihan data, representasi, algoritme, parameter, metrik, dan bentuk visualisasi; menjelaskan alternatif yang ditolak; serta menyatakan kondisi yang dapat membatalkan kesimpulan. Dokumen keputusan tersebut dapat dimasukkan ke portofolio dan menjadi dasar penyusunan proyek buku mini, tugas kelompok, skripsi, atau penelitian kolaboratif yang menghubungkan informatika kimia dengan kimia analitik, organik, anorganik, fisik, material, lingkungan, maupun farmasi.

Komponen audit	Pertanyaan kunci	Bukti minimum
career	Bagaimana career didefinisikan dan direpresentasikan?	Definisi, contoh positif-negatif, serta aturan transformasi.
lifelong learning	Bagaimana lifelong learning memengaruhi hasil atau keputusan?	Analisis sensitivitas, contoh struktur, dan catatan keterbatasan.
Kualitas data	Apakah sumber, unit, versi, dan provenance dapat ditelusuri?	Data dictionary, log kurasi, dan ringkasan sebelum-sesudah.
Validasi	Apakah evaluasi meniru skenario penggunaan nyata?	Split terdokumentasi, baseline, metrik, dan diagnosis error.
Reproduksibilitas	Dapatkah kelompok lain menjalankan ulang alur kerja?	Kode, environment, README, seed, parameter, dan checksum.

Praktikum 10 — Paket Riset Reprodusibel

Menyusun repositori, requirements, konfigurasi, log, uji sederhana, README, data card, dan arsip hasil.

Tujuan

- Mengubah konsep bab menjadi prosedur komputasi yang dapat diperiksa.
- Mencatat data, parameter, versi perangkat, dan keputusan analisis.
- Menyajikan hasil dalam tabel, gambar, dan narasi kimia yang konsisten.
- Mengidentifikasi keterbatasan serta rencana validasi lanjutan.

Bahan dan Perangkat

- Komputer dengan Python 3.11 atau lebih baru.
- Jupyter Notebook atau editor kode.
- RDKit, pandas, NumPy, matplotlib, dan scikit-learn sesuai kebutuhan.
- Dataset kecil yang legal digunakan dan dicatat sumber serta lisensinya.

Langkah Kerja

Buat folder proyek dan salin data mentah ke folder yang tidak diubah.
Tuliskan pertanyaan kimia, endpoint, unit, dan kriteria keberhasilan.
Audit struktur dan metadata; catat rekaman yang invalid atau meragukan.
Jalankan prosedur analisis tahap demi tahap; simpan parameter dan log.
Buat minimal dua visualisasi diagnostik dan satu tabel ringkas.
Lakukan pemeriksaan manual pada sedikitnya sepuluh rekaman.
Simpulkan hasil, ketidakpastian, domain aplikasi, dan pekerjaan lanjutan.
Kemas data terproses, kode, environment, README, dan laporan.

Contoh kode awal (perlu disesuaikan dengan dataset):

```
# Struktur proyek yang disarankan
project/
├── data_raw/          # tidak diubah
├── data_processed/    # hasil kurasi
├── notebooks/         # eksplorasi
├── src/               # fungsi/pipeline
├── tests/             # pengujian
├── environment.yml    # versi dependensi
├── README.md           # cara menjalankan
└── reports/            # tabel, gambar, dan naskah
```

Luaran Wajib

- Notebook atau skrip yang dapat dijalankan ulang dari awal.
- Data dictionary dan catatan kurasi.

- Tabel hasil utama serta minimal dua gambar dengan caption.
- Laporan 4–8 halaman yang memuat metode, hasil, pembahasan, batasan, dan kontribusi anggota.

Pertanyaan Analisis

- Keputusan apa yang berubah jika parameter utama diubah?
- Bagian mana yang paling rentan terhadap kesalahan kimia dan kesalahan komputasi?
- Apakah hasil berada dalam domain data dan metode yang digunakan?
- Bukti apa yang masih diperlukan sebelum hasil dipakai untuk eksperimen?

Ringkasan Bab

- Perangkat Lunak dan Masa Depan harus dipahami sebagai hubungan antara objek kimia, data, representasi, metode, dan keputusan.
- Kurasi, metadata, provenance, validasi, dan dokumentasi bukan tahap tambahan, melainkan bagian inti analisis.
- Metode sederhana yang transparan merupakan baseline wajib sebelum metode kompleks digunakan.
- Keluaran komputasi perlu dibaca bersama domain aplikasi, ketidakpastian, dan kebutuhan verifikasi eksperimen.
- Artefak kerja yang baik mencakup data, kode, parameter, environment, hasil, interpretasi, dan batasan.

Latihan dan Tugas Individu

1. Jelaskan dengan contoh hubungan antara toolkit dan cloud pada konteks menggabungkan editor molekul, RDKit, PostgreSQL, dan notebook.
2. Identifikasi minimal tiga risiko jika peneliti memilih paket tunggal untuk seluruh masalah tanpa evaluasi interoperabilitas.
3. Jelaskan dengan contoh hubungan antara open source dan vendor lock-in pada konteks prototipe dengan open source lalu deployment dengan infrastruktur terkelola.
4. Identifikasi minimal tiga risiko jika peneliti mengabaikan biaya total, lisensi, dan keberlanjutan.
5. Jelaskan dengan contoh hubungan antara Python dan scikit-learn pada konteks menghitung deskriptor dan fingerprint dari SMILES.
6. Identifikasi minimal tiga risiko jika peneliti notebook dengan sel dijalankan tidak berurutan dan dependensi tidak tercatat.
7. Jelaskan dengan contoh hubungan antara pipeline dan testing pada konteks menjalankan analisis ulang otomatis ketika dataset berubah.

8. Identifikasi minimal tiga risiko jika peneliti menyalin file manual antar-folder tanpa versioning.
9. Jelaskan dengan contoh hubungan antara Git dan data versioning pada konteks mencatat versi RDKit dan dataset dalam laporan.
10. Identifikasi minimal tiga risiko jika peneliti menganggap kode saja cukup untuk reproduksi.
11. Jelaskan dengan contoh hubungan antara unit test dan validation pada konteks uji bahwa standardisasi tidak mengubah jumlah atom berat.
12. Identifikasi minimal tiga risiko jika peneliti mengubah kode produksi tanpa pengujian regresi.
13. Jelaskan dengan contoh hubungan antara foundation model dan human oversight pada konteks model membaca literatur dan menghubungkan struktur dengan kondisi reaksi.
14. Identifikasi minimal tiga risiko jika peneliti menggunakan keluaran generatif tanpa pemeriksaan struktur dan referensi.
15. Jelaskan dengan contoh hubungan antara open science dan reproducibility pada konteks arsip dataset dan kode dengan DOI versi tetap.
16. Identifikasi minimal tiga risiko jika peneliti membagikan data tanpa metadata atau lisensi.
17. Jelaskan dengan contoh hubungan antara career dan lifelong learning pada konteks portofolio proyek QSAR yang mencakup data, kode, laporan, dan refleksi.
18. Identifikasi minimal tiga risiko jika peneliti mengukur kompetensi hanya dari jumlah perangkat yang pernah dipakai.
19. Buat peta konsep yang menghubungkan seluruh subbab dan beri satu contoh aliran data dari awal hingga keputusan.
20. Pilih satu artikel atau dataset terbuka, lalu audit representasi, provenance, metode, validasi, dan keterbatasannya.
21. Rancang eksperimen komputasi kecil yang dapat diselesaikan dalam dua minggu oleh kelompok tiga mahasiswa.
22. Tuliskan refleksi 300 kata: bagian mana yang paling membutuhkan pengetahuan kimia dan bagian mana yang paling membutuhkan pengetahuan komputasi?

Refleksi Bab

Tuliskan refleksi 500–700 kata yang menjawab: konsep apa yang paling mengubah cara Anda melihat molekul sebagai data; keputusan apa yang paling rentan terhadap kesalahan; serta bukti apa yang diperlukan sebelum keluaran komputasi digunakan untuk merancang eksperimen.

LAMPIRAN A. PANDUAN INSTALASI DAN LINGKUNGAN KOMPUTASI

Lingkungan belajar yang direkomendasikan menggunakan Python, conda atau mamba, JupyterLab, RDKit, pandas, NumPy, matplotlib, scikit-learn, dan paket pendukung. Versi perlu dicatat karena hasil standardisasi, descriptor, atau algoritme dapat berubah antarversi.

Contoh perintah instalasi:

```
# Contoh dengan conda/mamba
conda create -n informatika-kimia python=3.11
conda activate informatika-kimia
conda install -c conda-forge rdkit pandas numpy matplotlib scikit-learn
jupyterlab
python -c "import rdkit; print(rdkit.__version__)"
```

- Simpan file environment.yml atau requirements.txt.
- Catat sistem operasi, arsitektur, versi Python, RDKit, dan pustaka utama.
- Gunakan kernel Jupyter khusus lingkungan tersebut.
- Jalankan notebook dari awal setelah membersihkan output untuk menguji urutan eksekusi.
- Hindari memasang paket acak ke environment utama; buat lingkungan per proyek.

LAMPIRAN B. RINGKASAN FORMAT DAN IDENTIFIER KIMIA

Format/identifier	Isi utama	Kelebihan	Perhatian
SMILES	Notasi linear konektivitas dan stereokimia	Ringkas dan mudah diproses	Tidak selalu kanonik; model aromatisitas dapat berbeda.
SMARTS	Pola kueri substruktur	Ekspresif untuk pencarian dan transformasi	Kueri terlalu umum dapat menghasilkan banyak match.
InChI	Identifier berlapis berbasis struktur	Nonproprietary dan berguna untuk integrasi	Periksa versi serta standard/non-standard InChI.
InChIKey	Hash ringkas dari InChI	Mudah untuk pencocokan dan indeks web	Tidak dapat dibalik menjadi struktur lengkap.
MOL	Satu struktur dengan tabel atom/ikatan	Menyimpan koordinat dan atribut	Versi format dan dukungan stereokimia perlu diperiksa.
SDF	Banyak MOL plus bidang properti	Cocok untuk dataset molekul	Nama bidang dan tipe nilai tidak terstandar otomatis.
RXN/RD	Representasi reaksi	Memisahkan reaktan dan produk	Peran agen, kondisi, dan atom mapping perlu dilengkapi.
CSV/TSV	Tabel teks	Mudah dibaca berbagai perangkat	Struktur panjang, separator, encoding, dan tipe data rawan rusak.
JSON	Data hierarkis	Cocok untuk API dan metadata	Perlu skema untuk konsistensi.
RDF	Triple semantik	Integrasi data tertaut	Membutuhkan ontologi dan desain URI.

LAMPIRAN C. KARTU METODE DAN METRIK

Metode	Input	Penggunaan	Batasan utama
Tanimoto	Fingerprint biner	Kemiripan lokal/analog	Sangat bergantung pada fingerprint dan threshold.
PCA	Fitur kontinu yang diskalakan	Reduksi dimensi linear	Varians tinggi tidak selalu berarti relevansi endpoint.
Random forest	Tabular/fingerprint	Baseline nonlinear kuat	Probabilitas dapat tidak terkalibrasi; interpretasi global terbatas.
SVM	Fitur atau kernel	Klasifikasi/regresi ukuran sedang	Parameter dan scaling penting.
GNN	Graf molekul	Representasi terpelajar	Membutuhkan baseline, split ketat, dan evaluasi OOD.
Docking	Target 3D dan ligan	Prediksi pose/ranking	Skor bukan energi bebas; preparasi sistem dominan.
Nested CV	Dataset model	Estimasi performa setelah tuning	Biaya komputasi lebih tinggi, tetapi mengurangi optimisme.
Scaffold split	Struktur molekul	Uji generalisasi kimia	Ukuran set dan distribusi endpoint dapat tidak seimbang.
Conformal prediction	Prediksi dengan nonconformity score	Interval/set prediksi	Jaminan bergantung pada asumsi pertukaran data.

LAMPIRAN D. SEPULUH PROYEK BUKU MINI MAHASISWA

Setiap proyek dikerjakan oleh satu kelompok yang terdiri atas paling banyak tiga mahasiswa. Produk akhir berupa buku mini 60–100 halaman A4, file DOCX dan PDF, data/kode pendukung, serta presentasi. Buku mini wajib memuat kata pengantar, capaian pembelajaran, minimal lima bab, ilustrasi orisinal, studi kasus, praktikum, latihan, glosarium, daftar pustaka, kontribusi anggota, dan pernyataan penggunaan AI bila ada.

D.1 Bahasa Digital Molekul: Dari Graf Kimia ke Identitas Struktur

Cakupan utama: Bab 1–2. Buku mini yang menjelaskan representasi graf, format MOL/SDF, SMILES, SMARTS, InChI, stereokimia, dan pencarian substruktur dengan studi kasus senyawa obat atau bahan alam.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.2 Dunia Tiga Dimensi Molekul: Konformer, Farmakofor, dan Visualisasi

Cakupan utama: Bab 3. Buku mini berisi teori geometri, energi konformasi, algoritme konformer, alignment, RMSD, farmakofor, dan galeri visualisasi orisinal.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.3 Arsitektur Basis Data Kimia yang Bersih dan Terhubung

Cakupan utama: Bab 4. Buku mini yang merancang basis data molekul-assay, aturan standardisasi, provenance, API, serta contoh kueri SQL dan skema semantik.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.

- Berikan kritik metode, batasan, dan implikasi etis.

D.4 Atlas Ruang Kimia: Deskriptor, Fingerprint, dan Reduksi Dimensi

Cakupan utama: Bab 5. Buku mini yang membandingkan keluarga deskriptor, fingerprint, scaling, PCA/UMAP, dan interpretasi peta ruang kimia.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.5 Mengukur Kedekatan Molekul: Kemiripan, Activity Cliff, dan Keragaman

Cakupan utama: Bab 6. Buku mini yang membahas metrik, clustering, MCS, scaffold split, serta pemilihan subset beragam untuk eksperimen.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.6 QSAR yang Dapat Dipercaya: Dari Endpoint ke Model Card

Cakupan utama: Bab 7. Buku mini yang menyajikan baseline, machine learning, nested validation, metrik, domain aplikasi, ketidakpastian, dan interpretabilitas.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.7 Penapisan Virtual yang Bertanggung Jawab: Dari HTS ke Kandidat Eksperimen

Cakupan utama: Bab 8. Buku mini yang mengintegrasikan QC HTS, filter, ligand-based screening, docking, scoring, ADMET, dan validasi prospektif.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.8 Merancang Pustaka Senyawa: Enumerasi, Pareto, dan Retrosintesis

Cakupan utama: Bab 9. Buku mini yang membahas kimia kombinatorial, reaction SMARTS, desain multiobjektif, informatika reaksi, dan evaluasi kelayakan sintesis.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.9 Laboratorium Kimia Digital: ELN, LIMS, Otomasi, dan Siklus DMTA

Cakupan utama: Bab 9–10. Buku mini berupa rancangan laboratorium terhubung, model data, SOP digital, audit trail, dan skenario closed-loop yang tetap diawasi manusia.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.
- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

D.10 Masa Depan Informatika Kimia Indonesia: Open Science, AI, dan Pendidikan

Cakupan utama: Bab 1–10. Buku mini sintesis seluruh materi yang menawarkan peta jalan kurikulum, infrastruktur, riset, etika, dan portofolio mahasiswa di Indonesia.

- Rumusan masalah dan audiens pembaca harus dinyatakan eksplisit.
- Gunakan sedikitnya 20 sumber akademik; minimal lima berasal dari lima tahun terakhir.

- Sertakan sedikitnya enam ilustrasi orisinal dan empat tabel sintesis.
- Masukkan satu dataset atau studi kasus yang dapat direproduksi.
- Berikan kritik metode, batasan, dan implikasi etis.

LAMPIRAN E. SEPULUH TUGAS KELOMPOK

Setiap tugas dikerjakan oleh kelompok maksimal tiga mahasiswa. Anggota wajib melampirkan pembagian kerja dan log kontribusi. Dataset, kode, serta laporan harus dapat dijalankan atau diperiksa oleh kelompok lain.

E.1 Audit kualitas 200 rekaman kimia

Pilih dataset terbuka; identifikasi duplikasi, struktur invalid, missing value, satuan tidak konsisten, dan provenance. Buat laporan serta dataset hasil kurasi.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.2 Kamus konversi format struktur

Uji konversi SMILES, SDF, MOL, dan InChI untuk sedikitnya 30 molekul; dokumentasikan informasi yang hilang atau berubah.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.3 Eksperimen konformer terkontrol

Bandingkan jumlah konformer, energi, dan RMSD untuk lima molekul fleksibel pada dua konfigurasi ETKDG.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.4 Prototipe basis data molekul-assay

Bangun SQLite dengan sedikitnya empat tabel, 100 molekul, 3 assay, dan 300 hasil; sertakan kueri analitik.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.5 Peta ruang kimia kelas senyawa

Hitung minimal 10 deskriptor untuk 300 senyawa; lakukan PCA dan satu metode nonlinear; jelaskan distorsi dan keterbatasan.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.6 Benchmark fingerprint dan metrik

Bandingkan dua fingerprint dan tiga koefisien kemiripan untuk pencarian analog; analisis stabilitas peringkat.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.7 Baseline QSAR dan validasi ketat

Bangun sedikitnya tiga baseline, nested CV, scaffold split, uncertainty sederhana, dan model card.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.8 Triase virtual multi-tahap

Gabungkan filter sifat, kemiripan, model prediksi, dan penilaian manual; laporkan alasan memilih 20 kandidat akhir.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.9 Pustaka kombinatorial 1.000 produk

Enumerasi produk, validasi struktur, hitung sifat, buat front Pareto, dan pilih 50 produk untuk ditinjau.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

E.10 Paket reproduksi penelitian

Ambil salah satu tugas sebelumnya; susun repositori, data card, environment, tests, README, dan arsip hasil dengan checksum.

- Luaran: data mentah dan terproses, notebook/skrip, README, dua visualisasi, tabel hasil, laporan 8–15 halaman, dan presentasi 10 menit.
- Validasi: pemeriksaan manual, baseline, analisis sensitivitas, dan catatan domain aplikasi.
- Penilaian: ketepatan kimia 25%, metodologi 25%, reprodusibilitas 20%, interpretasi 20%, kerja sama dan komunikasi 10%.

LAMPIRAN F. RUBRIK PENILAIAN PROYEK DAN TUGAS

Aspek	Bobot	Kriteria sangat baik
Ketepatan konsep kimia	20%	Istilah, struktur, mekanisme, satuan, dan interpretasi benar serta konsisten.
Kualitas data dan provenance	15%	Sumber, lisensi, versi, metadata, kurasi, dan perubahan terdokumentasi.
Metode dan validasi	20%	Metode sesuai pertanyaan; baseline, split, metrik, dan analisis sensitivitas tepat.
Reprodusibilitas	15%	Kode berjalan, environment terkunci, instruksi jelas, hasil dapat diulang.
Analisis dan argumentasi	15%	Kesimpulan ditopang bukti, ketidakpastian dan batasan dibahas.
Komunikasi ilmiah	10%	Struktur naskah, tabel, gambar, caption, dan sitasi konsisten.
Kolaborasi dan integritas	5%	Kontribusi adil, penggunaan AI transparan, tidak ada fabrikasi atau plagiarisme.

LAMPIRAN G. TEMPLATE LAPORAN DAN DATA CARD

- Judul, anggota, tanggal, versi, dan tautan repositori.
- Ringkasan pertanyaan kimia dan keputusan yang didukung.
- Dataset: sumber, lisensi, tanggal akses, kriteria inklusi, jumlah awal-akhir.
- Representasi dan standardisasi: format, aturan garam, tautomer, muatan, stereokimia.
- Metode: fitur, algoritme, parameter, split, seed, environment.
- Validasi: metrik, interval, baseline, sensitivitas, domain aplikasi.
- Hasil: tabel utama, gambar diagnostik, contoh struktur, error analysis.
- Pembahasan: interpretasi kimia, perbandingan literatur, ketidakpastian, batasan.
- Reproducibility: perintah eksekusi, checksum, struktur folder, versi data dan kode.
- Pernyataan penggunaan AI dan verifikasi manusia.

LAMPIRAN H. CHECKLIST QUALITY CONTROL

Area	Pertanyaan pemeriksaan
Struktur	Semua SMILES dapat diparse; valensi, muatan, aromatisitas, dan stereokimia diperiksa.
Identitas	Duplikasi, garam, solvat, tautomer, dan parent compound ditangani sesuai tujuan.
Endpoint	Definisi, unit, arah, transformasi, dan protokol konsisten.
Metadata	Sumber, tanggal, versi, lisensi, dan proses transformasi tersedia.
Split	Tidak ada kebocoran analog, batch, waktu, atau sumber antara train dan test.
Fitur	Scaling dan seleksi fitur dilakukan di dalam pipeline data latih.
Model	Baseline, tuning, seed, dan rentang hiperparameter dilaporkan.
Metrik	Metrik sesuai tujuan dan ketidakseimbangan; variasi antar-split dilaporkan.
Interpretasi	Klaim tidak melampaui domain aplikasi; contoh sukses dan gagal dibahas.
Arsip	Data, kode, environment, log, README, dan hasil disimpan dengan versi.

LAMPIRAN I. GLOSARIUM A–Z

A

- ADMET — absorpsi, distribusi, metabolisme, ekskresi, dan toksisitas.
- Activity cliff — pasangan molekul sangat serupa dengan perbedaan aktivitas besar.
- Applicability domain — wilayah ruang kimia tempat model dianggap dapat digunakan.
- Atom mapping — korespondensi atom reaktan dan produk dalam reaksi.

B

- Basis data kimia — sistem penyimpanan terstruktur untuk struktur, sifat, reaksi, atau bioaktivitas.
- BCUT — keluarga deskriptor yang berasal dari eigenvalue matriks konektivitas berbobot.
- Bioaktivitas — respons biologis senyawa pada target atau sistem uji.
- Bit collision — beberapa fitur terpetakan pada posisi bit fingerprint yang sama.

C

- Canonicalization — prosedur menghasilkan representasi konsisten untuk struktur yang sama.
- Chemical space — ruang abstrak tempat molekul diposisikan berdasarkan deskriptor.
- Clustering — pengelompokan objek berdasarkan kemiripan atau jarak.
- Conformer — salah satu susunan 3D yang dapat dicapai molekul tanpa memutus ikatan.

D

- Data leakage — informasi dari data uji memengaruhi pelatihan atau pemilihan model.
- Descriptor — nilai numerik atau vektor yang mengodekan aspek molekul.
- Diversity — cakupan keragaman suatu koleksi dalam ruang kimia.
- Docking — prediksi pose dan/atau peringkat interaksi ligan pada situs target.

E

- ECFP — extended-connectivity fingerprint, fingerprint melingkar berbasis lingkungan atom.
- ELN — electronic laboratory notebook.
- Embedding — representasi vektor terpelajar untuk molekul atau objek kimia.
- Enrichment factor — ukuran pengayaan hit pada bagian awal ranking.

F

- FAIR — findable, accessible, interoperable, reusable.
- Fingerprint — representasi vektor fitur, sering biner, untuk struktur molekul.
- Force field — model energi klasik berbasis fungsi dan parameter empiris.
- Front Pareto — himpunan solusi yang tidak dapat diperbaiki pada satu tujuan tanpa memperburuk tujuan lain.

G

- Graph — struktur matematis berupa simpul dan sisi.
- Graph neural network — jaringan saraf yang melakukan agregasi pesan pada graf.
- Grid — kisi ruang atau parameter yang dipakai dalam perhitungan dan pencarian.
- Generative model — model yang menghasilkan kandidat molekul atau reaksi baru.

H

- Hashing — pemetaan fitur ke indeks tetap, sering pada fingerprint.
- High-throughput screening — pengujian banyak senyawa secara cepat dan terotomasi.
- Hydrogen-bond acceptor — atom atau fitur yang dapat menerima ikatan hidrogen.
- Hyperparameter — parameter proses belajar yang ditentukan di luar fitting model.

I

- Identifier — label stabil untuk mengenali entitas.
- InChI — identifier struktur kimia nonproprietary dari IUPAC.
- InChIKey — bentuk hash ringkas dari InChI.

- Isomorphism — kesetaraan graf melalui pemetaan simpul yang mempertahankan sisi.

J

- Join — operasi menggabungkan tabel basis data berdasarkan kunci.
- Jupyter Notebook — lingkungan dokumen komputasi interaktif.
- Jaccard/Tanimoto — koefisien overlap yang umum pada fingerprint biner.
- Jejak audit — rekaman perubahan dan tindakan yang dapat ditelusuri.

K

- Kanonisasi — padanan Indonesia untuk canonicalization.
- Kernel — fungsi kemiripan yang memungkinkan metode seperti SVM bekerja pada ruang implisit.
- K-means — algoritme klastering berbasis centroid.
- Kueri — permintaan pencarian atau pengambilan data.

L

- Ligand-based screening — penapisan berdasarkan informasi ligan yang diketahui.
- LIMS — laboratory information management system.
- LogP — logaritma koefisien partisi bentuk netral antara oktanol dan air.
- Latent space — ruang representasi terpelajar berdimensi lebih rendah.

M

- MCS — maximum common substructure.
- Metadata — data yang menjelaskan konteks, sumber, metode, dan arti data.
- Molecular graph — graf atom dan ikatan.
- Model card — dokumentasi tujuan, data, evaluasi, batasan, dan penggunaan model.

N

- Nested cross-validation — validasi bersarang yang memisahkan tuning dan estimasi performa.
- Normalization — transformasi skala data atau desain skema basis data untuk mengurangi redundansi.
- Novelty — derajat kebaruan kandidat terhadap data referensi.

- Node — simpul pada graf, biasanya atom pada graf molekul.

O

- Ontology — spesifikasi formal konsep dan relasi.
- Out-of-distribution — objek yang berbeda secara substansial dari distribusi data latih.
- Overfitting — model mempelajari noise atau kekhususan data latih.
- Open science — praktik membuka metode, data, dan kode secara bertanggung jawab.

P

- PAINS — substruktur yang sering terkait interferensi assay dan perlu ditinjau.
- Pharmacophore — susunan spasial fitur yang dianggap penting untuk aktivitas.
- PCA — principal component analysis.
- Provenance — catatan asal-usul dan transformasi data.

Q

- QSAR — quantitative structure–activity relationship.
- QSPR — quantitative structure–property relationship.
- Q^2 — ukuran prediktivitas silang pada konteks tertentu.
- Query molecule — molekul atau pola yang digunakan sebagai kueri.

R

- RDF — resource description framework.
- Reaction SMARTS — pola transformasi reaksi berbasis SMARTS.
- Reproducibility — kemampuan memperoleh hasil konsisten dengan data dan prosedur yang sama.
- RMSD — root-mean-square deviation koordinat.

S

- Scaffold — kerangka inti molekul.
- Scaffold split — pemisahan data berdasarkan kerangka untuk menguji generalisasi.
- SELFIES — representasi string molekul yang dirancang robust secara sintaksis.
- SMILES — notasi linear struktur molekul.

T

- Tanimoto coefficient — ukuran kemiripan berbasis fitur bersama dan total gabungan.
- Tautomer — bentuk struktur yang saling berubah melalui perpindahan proton dan ikatan rangkap.
- TPSA — topological polar surface area.
- Tversky index — koefisien kemiripan asimetris yang dapat memberi bobot berbeda.

U

- UMAP — metode proyeksi nonlinear untuk visualisasi data berdimensi tinggi.
- Uncertainty — ketidakpastian prediksi atau pengukuran.
- Unit test — pengujian bagian kecil kode secara terisolasi.
- URI — uniform resource identifier.

V

- Validation — evaluasi sistematis kinerja dan keterbatasan metode.
- Valence — kapasitas dan pola ikatan atom.
- Virtual screening — prioritisasi komputasional senyawa sebelum pengujian.
- Version control — pencatatan perubahan berkas dan kode dari waktu ke waktu.

W

- Workflow — rangkaian tahap analisis yang terdefinisi.
- Web service — layanan programatik melalui jaringan.
- Wiener index — indeks topologi berbasis jumlah jarak pasangan simpul.
- Widow/orphan — baris paragraf tunggal yang terpisah pada awal atau akhir halaman.

X

- XlogP — estimasi komputasional logP dengan metode tertentu.
- XML — format markup terstruktur yang digunakan dalam pertukaran data.
- X-ray crystallography — metode eksperimen penentuan struktur 3D dari difraksi sinar-X.
- X matrix — matriks fitur atau deskriptor dalam pemodelan.

Y

- Yield — rendemen reaksi.
- Y-randomization — pengacakan endpoint untuk mendeteksi korelasi kebetulan.
- Y vector — vektor respons/endpoint dalam pemodelan.
- Y-scrambling — istilah lain untuk pengacakan respons.

Z

- ZINC — basis data senyawa yang digunakan untuk virtual screening.
- Z-score — nilai hasil standardisasi terhadap rerata dan simpangan baku.
- Z-prime factor — ukuran kualitas assay high-throughput.
- Zero-shot prediction — prediksi tanpa contoh berlabel langsung untuk tugas tertentu.

LAMPIRAN J. DAFTAR PUSTAKA

Daftar pustaka berikut menggabungkan sumber utama, rujukan klasik, standar, basis data, perangkat, dan literatur pelengkap hingga 2025–2026. Versi dokumentasi perangkat dan basis data perlu diperiksa kembali ketika buku digunakan.

Baell, J. B., & Holloway, G. A. (2010). New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS). Journal of Medicinal Chemistry, 53, 2719–2740.

Bajorath, J. (Ed.). (2004). Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery. Humana Press.

Bajusz, D., Rácz, A., & Héberger, K. (2015). Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? Journal of Cheminformatics, 7, 20.

Baskaran, S., et al. (2025). Reporting Chemical Data in the Environmental Sciences. ACS Environmental Au, 5, 444–456. https://doi.org/10.1021/acsenvironau.5c00034

Bemis, G. W., & Murcko, M. A. (1996). The Properties of Known Drugs. 1. Molecular Frameworks. Journal of Medicinal Chemistry, 39, 2887–2893.

Bender, A., & Cortés-Ciriano, I. (2021). Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 1. Drug Discovery Today, 26, 511–524.

Bender, A., & Cortés-Ciriano, I. (2021). Artificial Intelligence in Drug Discovery: What Is Realistic, What Are Illusions? Part 2. Drug Discovery Today, 26, 1040–1052.

Bissantz, C., Folkers, G., & Rognan, D. (2000). Protein-Based Virtual Screening of Chemical Databases. Journal of Medicinal Chemistry, 43, 4759–4767.

Brown, N., Fiscato, M., Segler, M. H. S., & Vaucher, A. C. (2019). GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of Chemical Information and Modeling, 59, 1096–1108.

Bunin, B. A., Bajorath, J., Siesel, B., & Morales, G. (2007). Chemoinformatics: Theory, Practice, and Products. Springer.

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. arXiv:2010.09885.

Chowdhury, J., et al. (2024). Invariant Molecular Representations for Heterogeneous Catalysis. Journal of Chemical Information and Modeling, 64, 327–339. https://doi.org/10.1021/acs.jcim.3c00594

Coley, C. W., et al. (2019). A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. Science, 365, eaax1566.

Davies, M., et al. (2015). ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. Nucleic Acids Research, 43(W1), W612–W620.

Fourches, D., Muratov, E., & Tropsha, A. (2010). Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling. Journal of Chemical Information and Modeling, 50, 1189–1204.

Free, S. M., & Wilson, J. W. (1964). A Mathematical Contribution to Structure-Activity Studies. Journal of Medicinal Chemistry, 7, 395–399.

Gao, W., & Coley, C. W. (2020). The Synthesizability of Molecules Proposed by Generative Models. Journal of Chemical Information and Modeling, 60, 5714–5723.

Gasteiger, J., & Engel, T. (Eds.). (2003). Chemoinformatics: A Textbook. Wiley-VCH.

Gaulton, A., et al. (2017). The ChEMBL Database in 2017. Nucleic Acids Research, 45(D1), D945–D954.

Gilmer, J., et al. (2017). Neural Message Passing for Quantum Chemistry. Proceedings of ICML, 1263–1272.

Gramatica, P. (2007). Principles of QSAR Models Validation: Internal and External. QSAR & Combinatorial Science, 26, 694–701.

Hansch, C., & Fujita, T. (1964). ρ-σ-π Analysis: A Method for the Correlation of Biological Activity and Chemical Structure. Journal of the American Chemical Society, 86, 1616–1626.

Hastings, J., et al. (2016). ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites. Nucleic Acids Research, 44(D1), D1214–D1219.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., & Tchekhovskoi, D. (2015). InChI, the IUPAC International Chemical Identifier. Journal of Cheminformatics, 7, 23.

Hu, W., et al. (2020). Open Graph Benchmark: Datasets for Machine Learning on Graphs. NeurIPS 2020.

Irwin, J. J., & Shoichet, B. K. (2005). ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. Journal of Chemical Information and Modeling, 45, 177–182.

IUPAC. (2025). International Chemical Identifier (InChI). Compendium of Chemical Terminology, 5th ed., online version 5.0.0.

Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug Discovery with Explainable Artificial Intelligence. Nature Machine Intelligence, 2, 573–584.

Johnson, M. A., & Maggiora, G. M. (Eds.). (1990). Concepts and Applications of Molecular Similarity. Wiley.

Kim, J., Chang, W., Ji, H., & Joung, I. (2024). Quantum-Informed Molecular Representation Learning Enhancing ADMET Property Prediction. Journal of Chemical Information and Modeling, 64, 5028–5040. https://doi.org/10.1021/acs.jcim.4c00772

Kim, S., et al. (2025). PubChem 2025 Update. Nucleic Acids Research, 53(D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059

Kitchen, D. B., et al. (2004). Docking and Scoring in Virtual Screening for Drug Discovery. Nature Reviews Drug Discovery, 3, 935–949.

Klicpera, J., Groß, J., & Günnemann, S. (2020). Directional Message Passing for Molecular Graphs. ICLR 2020.

Krenn, M., et al. (2020). Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. Machine Learning: Science and Technology, 1, 045024.

Landrum, G. (2026). RDKit: Open-Source Cheminformatics, documentation version 2026.03.x. RDKit project. https://www.rdkit.org

Leach, A. R., & Gillet, V. J. (2007). An Introduction to Chemoinformatics (Revised ed.). Springer.

Lipinski, C. A., et al. (2001). Experimental and Computational Approaches to Estimate Solubility and Permeability. Advanced Drug Delivery Reviews, 46, 3–26.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. NeurIPS 2017.

Mayr, A., et al. (2018). Large-Scale Comparison of Machine Learning Methods for Drug Target Prediction. Chemical Science, 9, 5441–5451.

OECD. (2014). Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship Models. OECD Series on Testing and Assessment.

O'Boyle, N. M., et al. (2011). Open Babel: An Open Chemical Toolbox. Journal of Cheminformatics, 3, 33. https://doi.org/10.1186/1758-2946-3-33

Polishchuk, P. G., Madzhidov, T. I., & Varnek, A. (2013). Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. Journal of Computer-Aided Molecular Design, 27, 675–679.

Renz, P., et al. (2019). On Failure Modes in Molecule Generation and Optimization. Drug Discovery Today: Technologies, 32–33, 55–63.

Reymond, J.-L. (2015). The Chemical Space Project. Accounts of Chemical Research, 48, 722–730.

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling, 50, 742–754. https://doi.org/10.1021/ci100050t

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE, 10, e0118432.

Sandve, G. K., et al. (2013). Ten Simple Rules for Reproducible Computational Research. PLoS Computational Biology, 9, e1003285.

Schwaller, P., et al. (2019). Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. ACS Central Science, 5, 1572–1583.

Schwaller, P., et al. (2021). Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks. Nature Machine Intelligence, 3, 144–152.

Schütt, K. T., et al. (2018). SchNet—A Deep Learning Architecture for Molecules and Materials. Journal of Chemical Physics, 148, 241722.

Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. Nature, 555, 604–610.

Sheridan, R. P. (2013). Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. Journal of Chemical Information and Modeling, 53, 783–790.

Shoichet, B. K. (2006). Screening in a Spirit Haunted World. Drug Discovery Today, 11, 607–615.

Stokes, J. M., et al. (2020). A Deep Learning Approach to Antibiotic Discovery. Cell, 180, 688–702.

Tetko, I. V., et al. (2017). Virtual Computational Chemistry Laboratory—Design and Description. Journal of Computer-Aided Molecular Design, 31, 233–245.

Todeschini, R., & Consonni, V. (2009). Molecular Descriptors for Chemoinformatics (2nd ed.). Wiley-VCH.

Toplak, M., et al. (2014). Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. Journal of Chemical Information and Modeling, 54, 431–441.

Triballeau, N., et al. (2005). Virtual Screening Workflow Development Guided by the Receiver Operating Characteristic Curve Approach. Journal of Medicinal Chemistry, 48, 2534–2547.

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics, 29, 476–488.

Truchon, J.-F., & Bayly, C. I. (2007). Evaluating Virtual Screening Methods: Good and Bad Metrics for the Early Recognition Problem. Journal of Chemical Information and Modeling, 47, 488–508.

Vamathevan, J., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. Nature Reviews Drug Discovery, 18, 463–477.

Varnek, A., & Baskin, I. (2012). Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis? Journal of Chemical Information and Modeling, 52, 1413–1437.

Veber, D. F., et al. (2002). Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. Journal of Medicinal Chemistry, 45, 2615–2623.

Walters, W. P., & Murcko, M. A. (2020). Assessing the Impact of Generative AI on Medicinal Chemistry. Nature Biotechnology, 38, 143–145.

Warren, G. L., et al. (2006). A Critical Assessment of Docking Programs and Scoring Functions. Journal of Medicinal Chemistry, 49, 5912–5931.

Wilkinson, M. D., et al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, 3, 160018.

Willett, P. (1987). Similarity and Clustering in Chemical Information Systems. Research Studies Press.

Willett, P., Barnard, J. M., & Downs, G. M. (1998). Chemical Similarity Searching. Journal of Chemical Information and Computer Sciences, 38, 983–996.

Wu, Z., et al. (2018). MoleculeNet: A Benchmark for Molecular Machine Learning. Chemical Science, 9, 513–530.

Xie, J., et al. (2024). Self-Supervised Contrastive Molecular Representation Learning with a Chemical Synthesis Knowledge Graph. Journal of Chemical Information and Modeling, 64, 1945–1954. https://doi.org/10.1021/acs.jcim.4c00157

Yang, K., et al. (2019). Analyzing Learned Molecular Representations for Property Prediction. Journal of Chemical Information and Modeling, 59, 3370–3388.

Zhou, S., et al. (2024). Can Pretrained Models Really Learn Better Molecular Representations for AI-Aided Drug Discovery? Journal of Chemical Information and Modeling, 64, 2921–2930. https://doi.org/10.1021/acs.jcim.3c01707

INDEKS TOPIK TERPILIH

A–H	I–P	Q–Z
0D	1D	2D
3D	acak	accuracy
activity cliff	ADMET	akseptor
alignment	ambiguity	annotasi
API	applicability domain	aromatik
aromatisitas	artifact	assay
atom attribution	atom mapping	atom typing
audit	audit trail	automation
availability	barrier	basis data konformer
basis data struktur	BCUT	bias
binary vector	binding site	bioaktivitas
bit collision	boosting	building block
cache	cakupan klaim	calibration
candidate set	career	CAS
cell-based	ChEBI	checksum
ChEMBL	chemical language model	closed loop
cloud	cluster split	coefficient
cofactor	columnar	commercial
communication	completeness	condition prediction
confidence interval	confirmation	conformal prediction
connection table	connectivity index	consensus
consistency	constraint	container
controlled vocabulary	correlation filter	cosine
cost	coverage	CPSA
cross-validation	curse of dimensionality	DAG
data biologis	data kimia	data versioning
database cartridge	DBSCAN	decision
decision context	decision tree	decoy
deduplikasi	dendrogram	dependency
Dice	dihedral	disconnection
distance geometry	diverse library	DMTA
docking	document store	DOI
donor	dose response	E-state
E/Z	ECFP	elastic net
ELN	embedding	endpoint
enrichment	ensemble	enumeration
environment	Euclidean	external set
external validation	FAIR	feature selection

A–H	I–P	Q–Z
field descriptor	fingerprint	fitness for purpose
fleksibilitas	flexibility	focused library
foreign key	foundation model	garam
generative model	Git	GNN
graph database	graph generation	grid
hashing	HBA	HBD
heatmap	hidrofobik	hierarchical
high-throughput	hit rate	HTS
human oversight	hyperparameter	identifier
identity	imputation	InChI
InChIKey	indeks	indeks 3D
informasi	integration test	integritas
interdisciplinary	interdisiplin	interoperabilitas
interpretability	interpretation	isomorfisme
iteration	join	JSON
Jupyter	k-medoids	k-NN
kandidat	kanonisasi	kappa
Kartesian	kemometrika	kendala jarak
keputusan	kernel	ketidakpastian
kiralitas	konformer	koordinat
korespondensi atom	kristalografi	kunci jarak
label atom	lasso	latency
leakage	leverage	license
lifelong learning	LIMS	linkage
Lipinski	lisensi	literasi digital
locality	log transform	logging
logistic regression	logP	MACCS
MAE	Mahalanobis	Manhattan
margin	Markush	max-min
MCC	MCS	medan
message passing	metabolism	metadata
metode	minimum energi	MLR
MMP	model card	model fisik
MOL	molecular weight	monomer-based
Monte Carlo	multi-criteria	multicollinearity
multimodal	multiobjective	nested CV
neural network	NMR	normalisasi
normalization	novelty	objective hacking
ontology	OOD	open science
open source	orde ikatan	pagination

A–H	I–P	Q–Z
PAINS	pandas	panjang ikatan
parent	Pareto	paten
path fingerprint	PCA	PCR
pemrograman	pengetahuan	permeability
permukaan	pertanyaan riset	pharmacophore
pipeline	plate effect	PLS
portfolio	pose	pretraining
primary key	probability	product prediction
product property	product-based	produk
projection	properti	prospective validation
protein preparation	protonasi	protonation
provenance	PubChem	pusat reaksi
Python	QSAR	quality
query	R-group	R/S
R2	racemat	random forest
random split	ranking	rate limit
RDF	RDKit	reaction fingerprint
reaction rule	reactive group	reagent
reaksi	reaktan	reduced graph
regularization	rendemen	rendering
replicate	repository	representasi
representasi stik	representative set	reproducibility
rescoring	residual	resolusi
REST	retrieval	retrosynthesis
ridge	RMSD	RMSE
route search	ruang konformasi	salt stripping
sampling	SAR	scaffold
scaffold hopping	scaffold split	scaling
scikit-learn	scoring function	screening
SDF	search algorithm	selectivity
SELFIES	semantic web	sensitivitas
SHAP	shape	shape overlap
similar property principle	similarity search	simpul
sisi	sistematik	skala populasi
skema	SMARTS	SMILES
solubility	SOM	spektrum
sphere exclusion	standardization	statistika
stereoisomer	struktur	subgraf
sudut	support	SVM
synthetic accessibility	t-SNE	tabel

A–H	I–P	Q–Z
Tanimoto	target class	task definition
tautomer	testing	time split
timeout	toolkit	topologi
toxicity	TPSA	trade-off
transformasi	transparansi	triage
triple	Tversky	UMAP
uncertainty	unit	unit test
utility	valensi	validasi
validation	variance	variance threshold
Veber	vektor fitur	vendor lock-in
verifikasi	visualisasi	visualization
warna atom	WHIM	Wiener
workflow	yield	Z-prime
ZINC		


computation
computing
- KASMUI -
Dosen Kimia, Komputasi, IT, AI UNNES
https://hisabmu.com/
https://kasmui.cloud/
Ketua PCM Gunungpati 2
Anggota Majelis Tabligh PDM Kota Semarang & PWM Jawa Tengah
Tim Pengembang Software KHGT MTT PP Muhammadiyah
Praktisi Ilmu Falak



KASMUI

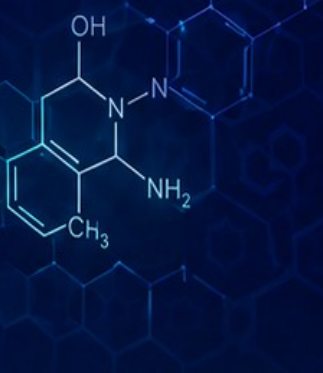
INFORMATIKA KIMIA

SINOPSIS

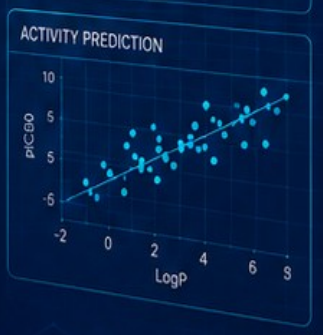
Buku INFORMATIKA KIMIA hadir sebagai panduan komprehensif bagi mahasiswa kimia S1 untuk memahami pertemuan antara ilmu kimia, data, dan komputasi. Buku ini membahas dasar-dasar representasi struktur kimia 2D dan 3D, basis data senyawa, deskriptor molekul, fingerprint, ruang kimia, serta konsep kemiripan dan keberagaman senyawa secara sistematis dan mudah dipahami.

Selain fondasi teori, buku ini juga menguraikan penerapan informatika kimia dalam analisis data, QSAR, machine learning, virtual screening, prediksi sifat molekul, docking, ADMET, dan desain pustaka senyawa. Setiap topik diperkuat dengan penjelasan praktis, ilustrasi, contoh alur kerja, dan orientasi pemecahan masalah yang relevan dengan dunia akademik maupun riset modern.

Dengan pendekatan yang terintegrasi, buku ini membantu pembaca membangun keterampilan berpikir komputasional dalam kimia, sekaligus menyiapkan mereka untuk proyek, praktikum, penulisan ilmiah, dan pengembangan riset berbasis data. INFORMATIKA KIMIA diharapkan menjadi rujukan penting bagi pembelajaran, penguatan literasi digital, dan penguasaan teknologi kimia masa kini.







KASMUI