

PAKET 15 PROYEK KERJA KELOMPOK

MATA KULIAH INFORMATIKA KIMIA

Proyek 4 Pekan dengan Produk Akhir Buku

A. KERANGKA UMUM PROYEK

1. Bentuk kegiatan

Setiap kelompok terdiri atas **maksimal 5 mahasiswa** dan mengerjakan satu proyek selama **4 pekan**.

Proyek tidak hanya berupa penulisan literatur. Setiap kelompok harus menghasilkan **produk akademik berbasis data**, yang menggabungkan:

1. kajian konsep informatika kimia;
2. pencarian dan kurasi data;
3. pengolahan atau analisis komputasi;
4. interpretasi berdasarkan prinsip kimia;
5. visualisasi;
6. validasi;
7. dokumentasi proses;
8. penyusunan hasil menjadi sebuah buku.

2. Produk akhir wajib

Produk utama:

BUKU INFORMATIKA KIMIA TEMATIK, 60–100 halaman A4.

Format:

- DOCX;
- PDF;
- minimal 5 bab;
- minimal 20 referensi akademik;
- sekurang-kurangnya 5 referensi dari lima tahun terakhir;
- minimal 6 gambar/diagram/visualisasi orisinal;
- minimal 4 tabel hasil atau tabel sintesis;
- minimal satu studi kasus berbasis dataset;
- satu praktikum/tutorial;
- soal latihan;

- glosarium;
- daftar pustaka;
- lampiran data;
- lampiran kode/notebook bila digunakan;
- log kontribusi setiap anggota;
- pernyataan penggunaan AI.

Produk pendukung:

- dataset mentah;
- dataset setelah kurasi;
- notebook Jupyter atau script Python;
- README;
- environment/package list;
- presentasi 10–15 menit;
- log pekerjaan kelompok.

B. CAPAIAN PEMBELAJARAN PROYEK

Setelah menyelesaikan proyek, mahasiswa diharapkan mampu:

1. menjelaskan molekul sebagai **objek kimia sekaligus objek informasi**;
2. memilih representasi molekul yang sesuai dengan tujuan analisis;
3. memperoleh dan mengkurasi data kimia secara bertanggung jawab;
4. menggunakan identifier seperti SMILES, InChI dan InChIKey;
5. menggunakan Python dan/atau RDKit pada persoalan informatika kimia;
6. menghitung atau memanfaatkan deskriptor dan fingerprint;
7. melakukan analisis kemiripan, clustering atau chemical space;
8. membangun atau mengevaluasi model informatika kimia;
9. melakukan validasi dan mengenali ketidakpastian;
10. menjelaskan keterbatasan prediksi komputasi;
11. menerapkan prinsip FAIR dan provenance;
12. menghasilkan analisis yang dapat direproduksi;
13. mengkomunikasikan hasil dalam bentuk **buku ilmiah mahasiswa**.

C. SILABUS PROYEK 4 PEKAN

Pkn.	Fokus	Aktivitas Pokok	Luaran
1	Perumusan masalah dan literatur	menentukan tema, pertanyaan proyek, mencari referensi, memilih dataset, menyusun outline buku	proposal 3–5 halaman, daftar pustaka awal, dataset awal, struktur buku
2	Data dan metode	kurasi data, representasi molekul, perhitungan/analisis awal, pembuatan kode, penulisan Bab 1–3	dataset terkurasi, notebook/script, 40–50% naskah
3	Analisis, validasi dan interpretasi	analisis utama, visualisasi, validasi, analisis sensitivitas/error, penulisan Bab 4–5	hasil final, tabel/gambar, 80% buku
4	Integrasi dan komunikasi ilmiah	revisi, pembahasan, praktikum, latihan, glosarium, referensi, proofreading, presentasi	buku DOCX/PDF, data, kode, README dan presentasi

Milestone

Akhir pekan 1: proposal disetujui dosen.

Akhir pekan 2: data dan metode harus dapat diperiksa.

Akhir pekan 3: hasil dan analisis selesai.

Akhir pekan 4: buku final dan presentasi.

D. PEMBAGIAN KERJA MAKSIMAL 5 MAHASISWA

Anggota 1 — Ketua proyek dan editor ilmiah

Mengkoordinasikan proyek, memastikan konsistensi konsep dan mengintegrasikan seluruh bab.

Anggota 2 — Kurator data dan literatur

Menangani dataset, provenance, metadata, referensi dan data dictionary.

Anggota 3 — Komputasi dan pemrograman

Menangani Python, RDKit, notebook, algoritme dan reproduksibilitas.

Anggota 4 — Analisis dan visualisasi

Menangani statistik, tabel, grafik, struktur molekul dan interpretasi.

Anggota 5 — Dokumentasi dan produksi buku

Menangani format buku, gambar, caption, daftar isi, glosarium, referensi dan presentasi.

Pembagian tersebut **bukan berarti setiap mahasiswa hanya menulis satu bagian**. Semua anggota tetap bertanggung jawab terhadap substansi keseluruhan buku.

E. LIMA BELAS PROYEK RINCI DAN SIAP DIKERJAKAN

PETUNJUK OPERASIONAL UNTUK SEMUA KELOMPOK

Seluruh proyek di bawah ini dirancang agar dapat dimulai pada hari pertama tanpa menunggu mahasiswa menentukan objek sendiri. Setiap proyek sudah memiliki objek/molekul awal, sumber data, ukuran dataset, kolom yang dipakai, perangkat lunak, langkah analisis, luaran mingguan, serta bentuk buku. Semua pekerjaan bersifat komputasi/digital; tidak diperlukan sintesis kimia atau penggunaan bahan kimia secara fisik.

Perangkat minimal yang digunakan

- Python, Jupyter Notebook, pandas, NumPy, matplotlib, scikit-learn, dan RDKit.
- Spreadsheet (Excel/LibreOffice Calc) untuk pemeriksaan manual dan dokumentasi.
- SQLite/DB Browser for SQLite untuk proyek basis data.
- Akses internet untuk mengunduh data dari sumber resmi pada awal proyek.
- Semua folder proyek harus memisahkan data mentah (raw) dari data terproses (processed). Data mentah tidak boleh ditimpa.

Sumber data utama yang dipakai berulang

Sumber resmi	Alamat	Dipakai untuk
PubChem (NIH)	Buka sumber resmi	nama, CID, SMILES, InChI, InChIKey, formula, sifat dasar
PubChem PUG REST	Buka sumber resmi	pengambilan data terprogram
COCONUT	Buka sumber resmi	dataset produk alam terbuka
QM9 / Figshare	Buka sumber resmi	struktur dan sifat kuantum 133.885 molekul kecil
ESOL / DeepChem	Buka sumber resmi	1128 struktur dengan log kelarutan air terukur
FreeSolv	Buka sumber resmi	energi bebas hidrasi molekul kecil netral

Sumber resmi	Alamat	Dipakai untuk
HOPV15	Buka sumber resmi	350 molekul/material fotovoltai organik dan sifat elektronik
Rhea	Buka sumber resmi	reaksi biokimia terkurasi dan ChEBI ID
SELFIES	Buka sumber resmi	representasi string molekul robust untuk proyek generatif

PROYEK 1

BAHASA DIGITAL MOLEKUL: REPRESENTASI 30 MOLEKUL NYATA DENGAN SMILES, InChI, InChIKey, SDF DAN SMARTS

Kaitan dengan materi kuliah

Bab 1–2: graf molekul, tabel konektivitas, format berkas, SMILES/SMARTS, InChI, stereokimia dan pencarian substruktur.

Pertanyaan proyek yang harus dijawab

Seberapa konsisten identitas kimia 30 molekul ketika direpresentasikan dan dikonversi antarformat digital?

Objek/dataset yang langsung digunakan

No.	Molekul tetap	Kategori
1	Air	molekul anorganik sederhana
2	Urea	senyawa karbonil sederhana
3	Glisin	asam amino
4	Alanin	asam amino
5	Valin	asam amino
6	Leusin	asam amino
7	Isoleusin	asam amino
8	Serin	asam amino
9	Treonin	asam amino

No.	Molekul tetap	Kategori
10	Asam aspartat	asam amino
11	Asam glutamat	asam amino
12	D-glukosa	karbohidrat
13	D-fruktosa	karbohidrat
14	Sukrosa	disakarida
15	Laktosa	disakarida
16	Maltosa	disakarida
17	Asam sitrat	asam organik
18	Asam malat	asam organik
19	Asam tartarat	asam organik
20	Asam laktat	asam organik
21	Asam askorbat	vitamin/antioksidan
22	Vanilin	produk alam/aroma
23	Eugenol	produk alam/aroma
24	Limonena	terpena
25	Mentol	terpena alkohol
26	Timol	fenolik alami
27	Sinamaldehida	produk alam/aroma
28	Kuersetin	flavonoid
29	Kurkumin	polifenol
30	Resveratrol	polifenol

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
PubChem Compound	Buka sumber resmi	ambil CID dan halaman ringkas untuk seluruh 30 nama

Sumber resmi	Alamat	Dipakai untuk
PUG REST Tutorial	Buka sumber resmi	otomasi pengambilan CanonicalSMILES, IsomericSMILES, InChI, InChIKey, MolecularFormula, MolecularWeight
RDKit Documentation	Buka sumber resmi	validasi struktur, konversi SDF/SMILES, SMARTS dan gambar 2D

Cara memperoleh data dan memulai pada hari pertama

- Buat file names.txt berisi 30 nama molekul pada tabel objek, satu nama per baris.
- Untuk uji pertama, buka PubChem PUG REST Tutorial dan ambil properti vanilin; setelah pola berhasil, jalankan untuk seluruh 30 nama.
- Buat notebook 00_download_pubchem.ipynb yang menyimpan respons mentah sebelum transformasi menjadi CSV.
- Instal lingkungan minimal: conda install -c conda-forge rdkit pandas jupyter matplotlib atau gunakan lingkungan RDKit yang sudah disediakan dosen.

Berkas yang harus dibuat kelompok

- molekul_30_raw.csv
- molekul_30_clean.csv
- molekul_30.sdf
- smart_query.csv
- notebook_p1.ipynb

Kolom/variabel minimum

- name
- PubChem CID
- canonical SMILES
- isomeric SMILES
- InChI
- InChIKey
- molecular formula
- molecular weight
- formal charge
- jumlah atom berat
- status parse RDKit

Prosedur rinci

1. Buat file daftar 30 nama molekul persis seperti tabel objek. Jangan mengganti daftar tanpa persetujuan dosen.

2. Ambil CID PubChem masing-masing nama. Jika nama memberikan lebih dari satu hasil, catat keputusan pemilihan CID dan alasannya.
3. Ambil Canonical SMILES, Isomeric SMILES, InChI, InChIKey, formula dan massa molekul dari PubChem melalui PUG REST atau ekspor CSV.
4. Baca semua SMILES dengan RDKit dan buat kolom parse_ok. Struktur yang gagal harus dicatat, bukan dihapus diam-diam.
5. Kanonisasi kembali SMILES dengan RDKit dan bandingkan dengan Canonical SMILES dari PubChem.
6. Tulis 30 struktur menjadi satu berkas SDF; baca ulang SDF dan hitung kembali InChIKey.
7. Uji round-trip SMILES → RDKit Mol → SDF → RDKit Mol → SMILES. Tandai apakah konektivitas dan stereokimia tetap konsisten.
8. Buat minimal 10 SMARTS untuk gugus fungsi: hidroksil, karboksilat/asam karboksilat, karbonil, amina, eter, aromatik, fenol, ikatan rangkap C=C, gula/poliol sederhana, dan cincin aromatik.
9. Jalankan 10 SMARTS tersebut terhadap seluruh 30 molekul dan buat matriks molekul × gugus fungsi.
10. Pilih minimal 5 kasus yang paling menarik untuk dibahas, misalnya perbedaan canonical/isomeric SMILES atau keberadaan banyak pusat kiral.
11. Buat atlas 2D 30 molekul dengan label nama, CID, dan satu identifier utama.
12. Susun rekomendasi: kapan sebaiknya memakai SMILES, InChI, InChIKey, dan SDF.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Ambil 30 CID dan seluruh identifier; verifikasi manual minimal 10 molekul.	raw CSV, daftar CID, 10 screenshot/rekaman verifikasi, outline Bab 1–2
2	RDKit parse, kanonisasi, ekspor/impor SDF, round-trip test.	clean CSV, SDF, notebook yang berjalan, tabel konsistensi
3	SMARTS search, matriks gugus fungsi, atlas 2D, analisis 5 kasus.	10 SMARTS, heatmap/matriks, 6+ gambar, draft Bab 3–5
4	Integrasi buku, praktikum langkah demi langkah, latihan dan glosarium.	DOCX/PDF 60–100 halaman + data + kode + presentasi

Gambar/tabel analisis minimum yang wajib muncul

- Atlas 2D 30 molekul
- diagram alir konversi format
- contoh round-trip sukses/gagal

- heatmap SMARTS vs molekul

Struktur minimum buku proyek

- Bab 1. Molekul sebagai objek informasi
- Bab 2. SMILES, InChI, InChIKey dan SDF
- Bab 3. Dataset 30 molekul dan prosedur kurasi
- Bab 4. Hasil konversi, SMARTS dan studi kasus
- Bab 5. Validasi, batasan dan rekomendasi format

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 2

DUNIA TIGA DIMENSI MOLEKUL: EKSPERIMEN KONFORMER 15 MOLEKUL DENGAN RDKit

Kaitan dengan materi kuliah

Bab 3: geometri 3D, ruang konformasi, pencarian konformer, energi, RMSD dan visualisasi.

Pertanyaan proyek yang harus dijawab

Bagaimana fleksibilitas struktur memengaruhi jumlah konformer, energi relatif dan keragaman bentuk molekul?

Objek/dataset yang langsung digunakan

No.	Molekul tetap	Alasan dipilih
1	Asam sitrat	fleksibel; 3 gugus karboksilat
2	Asam malat	fleksibel; 1 pusat kiral

No.	Molekul tetap	Alasan dipilih
3	Asam tartarat	dua pusat kiral
4	D-glukosa	banyak gugus OH
5	D-fruktosa	gula ketosa
6	Sukrosa	disakarida fleksibel
7	Vanilin	aromatik relatif kaku
8	Eugenol	aromatik dengan rantai alkenil
9	Limonena	cincin terpena
10	Mentol	sikloheksanol tersubstitusi
11	Linalool	rantai terpena fleksibel
12	Geraniol	rantai terpena fleksibel
13	Sinamaldehida	terkonjugasi relatif kaku
14	Kuersetin	flavonoid relatif kaku
15	Katekin	flavonoid 3D dengan beberapa pusat kiral

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
PubChem	Buka sumber resmi	SMILES awal 15 molekul
RDKit	Buka sumber resmi	ETKDG, MMFF/UFF, RMSD dan 3D conformer

Cara memperoleh data dan memulai pada hari pertama

- Gunakan file molekul_30_clean.csv dari Proyek 1 bila tersedia; ambil hanya 15 nama pada tabel Proyek 2.
- Buat notebook pertama yang hanya menguji satu molekul (vanilin): MolFromSmiles → AddHs → EmbedMultipleConfs(50, randomSeed=2026) → optimasi.
- Setelah satu molekul berhasil, jadikan fungsi reusable dan jalankan loop untuk 15 molekul.

Berkas yang harus dibuat kelompok

- konformer_input.csv
- konformer_all.sdf
- energi_rmsd.csv
- notebook_p2.ipynb

Kolom/variabel minimum

- name
- SMILES
- rotatable bonds
- number of generated conformers
- optimization status
- relative energy
- RMSD
- cluster conformer

Prosedur rinci

1. Ambil satu struktur/SMILES terverifikasi untuk masing-masing 15 molekul.
2. Buat molekul RDKit dan tambahkan hidrogen eksplisit.
3. Gunakan ETKDG untuk menghasilkan 50 konformer per molekul dengan randomSeed = 2026.
4. Optimalkan setiap konformer menggunakan MMFF94 bila parameter tersedia; gunakan UFF hanya untuk kasus yang tidak didukung dan catat alasannya.
5. Simpan energi setiap konformer dan ubah menjadi energi relatif terhadap konformer terendah.
6. Urutkan konformer; simpan 10 konformer berenergi terendah per molekul ke SDF.
7. Hitung RMSD berpasangan untuk 10 konformer tersebut.
8. Kelompokkan konformer secara sederhana berdasarkan ambang RMSD yang ditentukan dan dijelaskan kelompok.
9. Hitung jumlah rotatable bonds dan bandingkan dengan jumlah klaster konformer.
10. Buat plot energi relatif vs nomor konformer untuk setiap molekul dan ringkasan boxplot antar-molekul.
11. Pilih empat molekul: satu kaku, dua sedang, satu sangat fleksibel; buat galeri 3D dan jelaskan perbedaannya.
12. Buat tabel kegagalan optimasi atau anomali bila ada; jangan menghapus kasus gagal tanpa pelaporan.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Verifikasi 15 struktur dan buat pipeline ETKDG pada 3 molekul uji.	input CSV + notebook prototipe
2	Jalankan 50 konformer × 15 molekul; optimasi dan simpan energi.	SDF semua konformer + tabel energi
3	RMSD, clustering, hubungan fleksibilitas-konformer dan visualisasi.	energi_rmsd.csv + grafik + galeri 3D
4	Tulis buku dan praktikum “Generasi dan Analisis Konformer”.	buku final + kode + presentasi

Gambar/tabel analisis minimum yang wajib muncul

- diagram workflow ETKDG
- plot energi konformer
- matriks RMSD 4 molekul
- galeri konformer kaku vs fleksibel

Struktur minimum buku proyek

- Bab 1. Struktur 3D dan ruang konformasi
- Bab 2. ETKDG dan optimasi force field
- Bab 3. Desain eksperimen 15 molekul
- Bab 4. Energi, RMSD dan clustering
- Bab 5. Interpretasi, keterbatasan dan reproduksibilitas

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 3

MEMBANGUN BASIS DATA KIMIA NYATA: 300 PRODUK ALAM COCONUT KE SQLITE

Kaitan dengan materi kuliah

Bab 4: basis data, standardisasi, deduplikasi, kualitas data, provenance, API/interoperabilitas.

Pertanyaan proyek yang harus dijawab

Bagaimana 300 rekaman produk alam dapat diubah dari file datar menjadi basis data kimia yang konsisten, dapat ditelusuri dan dapat dikueri?

Objek/dataset yang langsung digunakan

Gunakan tepat 300 rekaman dari COCONUT. Ambil file CSV resmi, lalu filter: struktur dapat diparse RDKit, formal charge = 0 bila kolom tersedia, $100 \leq \text{molecular weight} \leq 600$, tidak mengandung atom logam, dan memiliki identifier/SMILES. Setelah filter, ambil sampel 300 baris

dengan `random_state = 2026` dan simpan daftar COCONUT ID agar kelompok lain dapat mereplikasi subset yang sama.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
COCONUT Download	Buka sumber resmi	CSV/SDF produk alam
COCONUT	Buka sumber resmi	verifikasi record dan metadata
RDKit	Buka sumber resmi	standardisasi dan descriptor

Cara memperoleh data dan memulai pada hari pertama

- Buka halaman COCONUT Download, unduh CSV rilis terbaru yang menyediakan ID dan SMILES, lalu simpan persis sebagai `data/raw/coconut_original.csv`.
- Buat notebook `01_select_300.ipynb`: baca CSV, deteksi otomatis nama kolom ID/SMILES, lakukan filter proyek, lalu `sample(n=300, random_state=2026)`.
- Simpan `selected_coconut_ids.txt` sebelum membuat database agar subset tidak berubah pada run berikutnya.

Berkas yang harus dibuat kelompok

- `coconut_raw.csv`
- `coconut_300_selected.csv`
- `coconut_300_clean.csv`
- `coconut_300.sqlite`
- `schema.sql`
- `queries.sql`
- `notebook_p3.ipynb`

Kolom/variabel minimum

- `coconut_id`
- `name`
- `smiles`
- `inchikey`
- `molecular formula`
- `MW`
- `NP-likeness/QED` bila tersedia
- `source/reference` bila tersedia
- `canonical_smiles`
- `Murcko scaffold`

Prosedur rinci

1. Unduh file COCONUT resmi dan simpan tanpa perubahan sebagai data mentah.

2. Baca dataset dengan pandas; catat jumlah awal dan nama kolom asli.
3. Parse SMILES memakai RDKit; buat log seluruh record gagal parse.
4. Terapkan kriteria filter proyek, kemudian ambil tepat 300 record dengan random_state=2026.
5. Kanonisasi SMILES dan hitung InChIKey dengan RDKit untuk kunci identitas tambahan.
6. Hitung descriptor minimal: MolWt, LogP, TPSA, HBD, HBA, rotatable bonds, ring count, fraction Csp3.
7. Hitung Bemis–Murcko scaffold.
8. Rancang SQLite minimal empat tabel: molecule, descriptor, source/provenance, scaffold.
9. Tentukan primary key, foreign key, tipe data dan aturan null.
10. Impor 300 record ke SQLite.
11. Tulis minimal 12 query SQL: kisaran MW, top TPSA, jumlah scaffold unik, kelompok LogP, pencarian nama, pencarian InChIKey, dan query join.
12. Bandingkan hasil query SQL dengan perhitungan pandas untuk validasi silang.
13. Buat diagram ERD dan data dictionary.
14. Ekspor satu view hasil gabungan menjadi CSV untuk digunakan kelompok lain.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Unduh COCONUT, pahami kolom dan tetapkan 300 ID.	raw file + selected IDs + data dictionary awal
2	Kurasi, canonicalization, descriptor dan scaffold.	clean CSV + log kurasi
3	Bangun SQLite, 12 query, ERD, validasi silang.	SQLite + schema.sql + queries.sql + ERD
4	Buku panduan basis data dan demonstrasi query.	buku + database + notebook + presentasi

Gambar/tabel analisis minimum yang wajib muncul

- ERD basis data
- flowchart raw→clean→SQLite
- distribusi MW/LogP
- 10 scaffold paling sering

Struktur minimum buku proyek

- Bab 1. Sumber data produk alam dan provenance

- Bab 2. Kurasi dan standardisasi
- Bab 3. Desain basis data relasional
- Bab 4. Implementasi dan query SQL
- Bab 5. Audit kualitas dan interoperabilitas

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 4

AUDIT FAIR DAN REKONSILIASI IDENTITAS: 100 RECORD COCONUT DIBANDINGKAN DENGAN PUBCHEM

Kaitan dengan materi kuliah

Bab 1 dan 4: FAIR, provenance, identifier, rekonsiliasi identitas, kualitas data dan interoperabilitas.

Pertanyaan proyek yang harus dijawab

Apakah satu molekul yang sama memiliki identitas, nama, formula dan struktur digital yang konsisten ketika dibandingkan antara dua basis data terbuka?

Objek/dataset yang langsung digunakan

Ambil 100 record pertama dari file coconut_300_selected.csv milik Proyek 3 (atau buat subset yang sama dengan aturan tersebut). Untuk setiap record, gunakan InChIKey sebagai kunci utama untuk mencari record yang sesuai di PubChem. Jika tidak ditemukan, gunakan canonical SMILES sebagai pencarian kedua dan catat metode pencocokannya.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
COCONUT	Buka sumber resmi	sumber identitas A
PubChem	Buka sumber resmi	sumber identitas B
PubChem PUG REST	Buka sumber resmi	pencarian/ekstraksi PubChem secara terprogram

Cara memperoleh data dan memulai pada hari pertama

- Gunakan 100 baris pertama coconut_300_selected.csv; bila Proyek 3 tidak ada di kelas, buat sendiri subset 300 dengan aturan yang sama lalu ambil 100 pertama.
- Mulai dengan 10 InChIKey sebagai pilot: cari PubChem CID dan properti, baru setelah logika matching benar jalankan 100 record.
- Buat kolom match_method sejak awal: InChIKey, canonical_SMILES, atau manual_review.

Berkas yang harus dibuat kelompok

- audit_100_input.csv
- pubchem_matches.csv
- identity_conflicts.csv
- fair_scorecard.xlsx
- notebook_p4.ipynb

Kolom/variabel minimum

- COCONUT ID
- COCONUT name
- COCONUT SMILES
- COCONUT InChIKey
- PubChem CID
- PubChem name
- PubChem canonical SMILES
- PubChem InChIKey
- formula A/B
- MW A/B
- match_status
- conflict_type

Prosedur rinci

1. Ambil 100 record COCONUT dan simpan sebagai input yang tidak berubah.
2. Cari PubChem CID menggunakan InChIKey; catat record yang tidak ditemukan.
3. Ambil nama, canonical SMILES, InChIKey, formula dan molecular weight dari PubChem.

4. Bandingkan InChIKey; klasifikasikan exact match, unmatched, atau ambiguous.
5. Bandingkan formula dan massa molekul dengan toleransi numerik yang jelas.
6. Bandingkan canonical SMILES setelah kedua SMILES dikanonisasi ulang menggunakan RDKit.
7. Buat tipe konflik: perbedaan nama/sinonim, perbedaan protonasi/muatan, perbedaan stereokimia, perbedaan garam/parent, record tidak ditemukan, atau konflik lain.
8. Hitung persentase record pada setiap tipe konflik.
9. Pilih 10 kasus konflik paling informatif dan lakukan pemeriksaan manual halaman sumber.
10. Buat scorecard FAIR sederhana untuk Findable, Accessible, Interoperable, Reusable; kriteria harus dijelaskan sebelum scoring.
11. Buat rekomendasi aturan identity reconciliation untuk proyek kimia berikutnya.
12. Simpan seluruh keputusan manual dalam kolom curator_note.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Tetapkan 100 input dan skrip matching PubChem.	input 100 + 30 match uji
2	Selesaikan 100 match dan klasifikasi konflik.	pubchem_matches.csv + conflicts.csv
3	FAIR scorecard, pemeriksaan manual 10 kasus dan rekomendasi.	scorecard + tabel konflik + gambar kasus
4	Buku audit FAIR dan tutorial rekonsiliasi identitas.	buku final + data + notebook

Gambar/tabel analisis minimum yang wajib muncul

- diagram exact/ambiguous/unmatched
- bar chart tipe konflik
- 10 contoh perbandingan struktur
- radar/bar FAIR score

Struktur minimum buku proyek

- Bab 1. FAIR dan identitas kimia
- Bab 2. Desain audit dua basis data
- Bab 3. Rekonsiliasi identifier dan struktur
- Bab 4. Hasil konflik dan FAIR score
- Bab 5. SOP rekonsiliasi dan keterbatasan

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.

- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 5

ATLAS RUANG KIMIA QM9: PCA DAN PETA 2.000 MOLEKUL KECIL

Kaitan dengan materi kuliah

Bab 5: deskriptor, scaling, korelasi, seleksi fitur, PCA dan visualisasi ruang kimia.

Pertanyaan proyek yang harus dijawab

Bagaimana 2.000 molekul kecil dapat dipetakan dalam ruang deskriptor, dan fitur apa yang paling menentukan pemisahan wilayah ruang kimia?

Objek/dataset yang langsung digunakan

Gunakan dataset QM9. Dari keseluruhan dataset, ambil 2.000 molekul secara acak dengan `random_state = 2026` setelah struktur dapat diparse. Jika menggunakan DeepChem/MoleculeNet, simpan ID/SMILES hasil sampling agar subset dapat direplikasi.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
QM9 Figshare	Buka sumber resmi	dataset utama
DeepChem	Buka sumber resmi	opsi loader QM9
MoleculeNet		
RDKit	Buka sumber resmi	descriptor dan fingerprint

Cara memperoleh data dan memulai pada hari pertama

- Cara termudah: gunakan DeepChem MoleculeNet `load_qm9` atau unduh koleksi QM9 Figshare; jangan melakukan komputasi DFT baru.
- Ekstrak ID/SMILES, kemudian sample 2.000 record dengan `random_state=2026` dan simpan `qm9_subset_ids.txt`.

- Uji descriptor dan PCA lebih dahulu pada 100 record; setelah pipeline stabil, jalankan 2.000.

Berkas yang harus dibuat kelompok

- qm9_subset_2000.csv
- qm9_descriptors.csv
- pca_scores.csv
- notebook_p5.ipynb

Kolom/variabel minimum

- SMILES/ID
- heavy atoms
- C/N/O/F count
- MolWt
- TPSA
- LogP
- HBD
- HBA
- rotatable bonds
- ring count
- fraction Csp3
- PCA1
- PCA2

Prosedur rinci

1. Muat QM9 dan ambil subset 2.000 record yang terdokumentasi.
2. Parse struktur dengan RDKit; pisahkan record gagal bila ada.
3. Hitung 8–12 descriptor sederhana yang tidak memerlukan komputasi kuantum baru.
4. Buat statistik deskriptif dan histogram tiap descriptor.
5. Hitung matriks korelasi; tandai pasangan descriptor dengan korelasi absolut tinggi.
6. Lakukan standardization menggunakan StandardScaler.
7. Lakukan PCA dan laporkan explained variance PC1–PC5.
8. Buat scatter PC1–PC2 dengan pengelompokan berdasarkan jumlah heavy atoms.
9. Buat scatter kedua dengan warna berdasarkan jumlah heteroatom (N+O+F).
10. Pilih 20 molekul ekstrem pada PC1/PC2 dan tampilkan struktur 2D-nya.

11. Bandingkan PCA berbasis descriptor dengan PCA berbasis subset fitur yang direduksi setelah korelasi tinggi dihapus.

12. Jelaskan secara kimia descriptor apa yang mendorong arah PC1 dan PC2.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Ambil 2.000 struktur, hitung descriptor dan QC.	subset CSV + descriptor CSV
2	Korelasi, scaling dan PCA awal.	correlation matrix + pca scores
3	Visualisasi, analisis molekul ekstrem dan perbandingan feature set.	4–6 grafik + atlas 20 struktur
4	Susun Atlas Ruang Kimia QM9.	buku + notebook + data

Gambar/tabel analisis minimum yang wajib muncul

- heatmap korelasi
- scree plot PCA
- PC1-PC2 menurut heavy atoms
- atlas 20 molekul ekstrem

Struktur minimum buku proyek

- Bab 1. Konsep ruang kimia
- Bab 2. QM9 dan deskriptor
- Bab 3. Scaling, korelasi dan PCA
- Bab 4. Atlas 2.000 molekul
- Bab 5. Interpretasi dan batas PCA

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 6

KEMIRIPAN MOLEKUL DAN “SOLUBILITY CLIFF”: ANALISIS 1.128 MOLEKUL ESOL

Kaitan dengan materi kuliah

Bab 6: fingerprint, Tanimoto, clustering, similarity principle dan activity/property cliff.

Pertanyaan proyek yang harus dijawab

Apakah molekul yang sangat mirip secara struktur selalu mempunyai kelarutan air yang juga mirip?

Objek/dataset yang langsung digunakan

Gunakan seluruh dataset ESOL/Delaney yang berisi 1.128 senyawa. Endpoint utama adalah measured log solubility in mols per litre. Definisi operasional property cliff proyek: pasangan dengan Tanimoto ECFP4 $\geq 0,80$ dan $|\Delta \log S| \geq 1,00$.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
ESOL CSV	Buka sumber resmi	dataset 1.128 senyawa
DeepChem tutorial	Buka sumber resmi	keterangan endpoint log kelarutan
RDKit	Buka sumber resmi	Morgan fingerprint dan similarity

Cara memperoleh data dan memulai pada hari pertama

- Unduh delaney-processed.csv (ESOL) dari repository DeepChem dan simpan sebagai data/raw/esol.csv.
- Pastikan jumlah record data adalah 1.128 setelah header dibaca; catat bila versi sumber berbeda.
- Uji fingerprint Morgan radius=2, nBits=2048 untuk 20 record pertama, lalu jalankan seluruh dataset.

Berkas yang harus dibuat kelompok

- esol_raw.csv
- esol_clean.csv
- similarity_pairs.csv
- solubility_cliffs.csv
- notebook_p6.ipynb

Kolom/variabel minimum

- Compound ID
- SMILES

- measured logS
- Morgan fingerprint
- nearest neighbor
- Tanimoto
- delta logS
- cliff flag

Prosedur rinci

1. Unduh ESOL CSV dan simpan versi mentah.
2. Validasi semua SMILES dengan RDKit dan buat canonical SMILES.
3. Hitung Morgan/ECFP fingerprint radius 2, nBits 2048.
4. Hitung nearest neighbor setiap molekul berdasarkan Tanimoto.
5. Buat distribusi nilai nearest-neighbor similarity.
6. Hitung $\Delta\log S$ terhadap nearest neighbor.
7. Terapkan definisi solubility cliff: similarity $\geq 0,80$ dan $|\Delta\log S| \geq 1,00$.
8. Buat tabel 20 cliff terkuat, diurutkan berdasarkan $|\Delta\log S|$ lalu similarity.
9. Gambar pasangan struktur cliff berdampingan beserta logS.
10. Lakukan clustering sederhana terhadap fingerprint dan periksa variasi logS di dalam cluster.
11. Bandingkan Tanimoto dengan Dice untuk 20 pasangan terpilih.
12. Simpulkan kapan similarity principle gagal pada data ini.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Download/QA ESOL dan fingerprint seluruh record.	clean CSV + fingerprint notebook
2	Nearest-neighbor similarity dan identifikasi cliff.	similarity_pairs + cliff table
3	Clustering, Tanimoto vs Dice, galeri 20 pasangan.	heatmap/dendrogram + galeri
4	Buku “Kemiripan tidak selalu berarti sifat sama”.	buku + data + kode

Gambar/tabel analisis minimum yang wajib muncul

- histogram nearest-neighbor similarity
- scatter similarity vs $|\Delta\log S|$
- galeri 20 cliff
- dendrogram/cluster plot

Struktur minimum buku proyek

- Bab 1. Similarity principle
- Bab 2. Fingerprint dan koefisien kemiripan
- Bab 3. ESOL dan desain property-cliff
- Bab 4. Hasil dan studi pasangan
- Bab 5. Keterbatasan dan implikasi

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 7

QSPR KELARUTAN AIR: MEMBANGUN DAN MEMVALIDASI MODEL ESOL DENGAN TIGA ALGORITME

Kaitan dengan materi kuliah

Bab 7: endpoint, feature engineering, regresi, train/test split, cross-validation, metrik, error analysis dan model card.

Pertanyaan proyek yang harus dijawab

Seberapa baik deskriptor molekuler sederhana dapat memprediksi log kelarutan air pada ESOL tanpa data leakage?

Objek/dataset yang langsung digunakan

Gunakan ESOL 1.128 senyawa. Endpoint: measured log solubility in mols per litre. Fitur wajib: MolWt, MolLogP, TPSA, HBD, HBA, rotatable bonds, ring count dan FractionCSP3. Model wajib: Linear Regression, Random Forest Regressor, dan k-Nearest Neighbors Regressor.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
ESOL CSV	Buka sumber resmi	struktur + endpoint
scikit-learn	Buka sumber resmi	pipeline, model dan metrik
RDKit	Buka sumber resmi	descriptor dan Murcko scaffold

Cara memperoleh data dan memulai pada hari pertama

- Gunakan file ESOL yang sama dengan Proyek 6 agar hasil antar-kelompok dapat dibandingkan.
- Buat satu file `split_assignment.csv` dan jangan mengubahnya setelah model pertama dijalankan.
- Jalankan baseline terlebih dahulu; model yang lebih kompleks baru dinilai bila mampu mengalahkan baseline secara konsisten.

Berkas yang harus dibuat kelompok

- `esol_qspr.csv`
- `split_assignment.csv`
- `model_results.csv`
- `error_cases.csv`
- `model_card.md`
- `notebook_p7.ipynb`

Kolom/variabel minimum

- SMILES
- endpoint logS
- 8 descriptors
- scaffold
- split
- prediction_LR
- prediction_RF
- prediction_KNN
- residuals

Prosedur rinci

1. Baca dan validasi ESOL; hitung delapan descriptor wajib.
2. Buat Murcko scaffold setiap molekul.
3. Buat split berbasis scaffold sehingga scaffold yang sama tidak tersebar bebas antara train dan test; simpan assignment ke CSV.
4. Gunakan sekitar 80% data untuk pengembangan dan 20% sebagai test eksternal; di dalam data pengembangan gunakan 5-fold cross-validation.

5. Bangun baseline berupa prediksi rerata endpoint data train.
6. Bangun Linear Regression dalam pipeline scaling.
7. Bangun k-NN dalam pipeline scaling; tuning hanya pada data pengembangan.
8. Bangun Random Forest dengan random_state=2026.
9. Laporkan MAE, RMSE dan R² pada cross-validation dan test.
10. Buat predicted-vs-observed dan residual plot untuk ketiga model.
11. Identifikasi 20 error terbesar pada model terbaik dan gambar strukturnya.
12. Bandingkan domain descriptor train vs test dan tandai test point yang berada di luar rentang beberapa descriptor.
13. Susun model card: tujuan, data, endpoint, fitur, split, metrik, keterbatasan, failure modes dan cara penggunaan yang tepat.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Descriptor + scaffold + split terkunci.	esol_qspr.csv + split_assignment.csv
2	Baseline + 3 model + CV.	notebook model + CV table
3	Test set, error analysis, applicability domain sederhana.	results + error_cases + grafik
4	Model card dan buku QSPR.	buku + model card + kode

Gambar/tabel analisis minimum yang wajib muncul

- predicted vs observed tiga model
- residual plots
- feature importance RF
- galeri 20 error terbesar

Struktur minimum buku proyek

- Bab 1. QSPR dan endpoint kelarutan
- Bab 2. Descriptor dan split berbasis scaffold
- Bab 3. Tiga algoritme dan validasi
- Bab 4. Hasil, error dan domain aplikasi
- Bab 5. Model card dan rekomendasi

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.

- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 8

PENAPISAN VIRTUAL BAHAN FOTOVOLTAIK ORGANIK: MERANKING 350 MOLEKUL HOPV15

Kaitan dengan materi kuliah

Bab 5, 7 dan 8: descriptor/fitur, ranking, validasi model, virtual screening dan keputusan multi-kriteria.

Pertanyaan proyek yang harus dijawab

Dapatkah sifat elektronik terhitung digunakan untuk meranking kandidat material fotovoltaiik organik dan seberapa baik ranking tersebut sesuai dengan PCE eksperimen?

Objek/dataset yang langsung digunakan

Gunakan seluruh 350 molekul HOPV15. Fokus hanya pada data komputasi dan eksperimen yang sudah tersedia; tidak melakukan sintesis atau eksperimen fisik. Gunakan HOMO, LUMO, HOMO-LUMO gap dan fitur numerik lain yang tersedia untuk memprediksi/ranking PCE eksperimen.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
HOPV15 paper/data descriptor	Buka sumber resmi	350 molekul, PCE, HOMO, LUMO, gap, koordinat
DeepChem	Buka sumber resmi	opsi loader HOPV
MoleculeNet		
RDKit	Buka sumber resmi	descriptor tambahan dari struktur

Cara memperoleh data dan memulai pada hari pertama

- Gunakan DeepChem `dc.molnet.load_hopv(...)` untuk memperoleh dataset HOPV secara langsung, atau unduh data dari halaman Scientific Data.

- Buat data dictionary yang memisahkan properti eksperimen (terutama PCE) dari fitur/hasil perhitungan.
- Pada hari pertama, cukup buat scatter HOMO vs LUMO dan histogram PCE untuk memastikan data terbaca benar.

Berkas yang harus dibuat kelompok

- hopv_raw.csv
- hopv_model_data.csv
- hopv_ranking.csv
- top20_candidates.csv
- notebook_p8.ipynb

Kolom/variabel minimum

- molecule ID/SMILES
- experimental PCE
- HOMO
- LUMO
- gap
- optional RDKit descriptors
- predicted PCE
- rank predicted
- rank experimental

Prosedur rinci

1. Ambil HOPV15 dan identifikasi kolom dengan PCE eksperimen serta sifat elektronik terhitung.
2. Buat tabel kelengkapan data; jangan imputasi PCE yang tidak tersedia tanpa alasan.
3. Gunakan hanya record dengan endpoint PCE eksperimen untuk membangun model evaluasi.
4. Buat fitur minimal HOMO, LUMO dan gap; tambahkan 3–5 descriptor RDKit bila struktur mudah tersedia.
5. Bagi data menjadi train/test dengan random_state=2026; simpan split.
6. Bangun baseline sederhana dan satu model regresi nonlinier, misalnya Random Forest.
7. Evaluasi MAE/RMSE/R² pada test.
8. Gunakan prediksi test untuk membuat ranking kandidat tanpa melihat PCE eksperimen saat ranking dibuat.
9. Bandingkan top-20 predicted dengan top-20 experimental; hitung overlap dan Spearman rank correlation.

10. Buat tabel kandidat yang over-ranked dan under-ranked.
11. Pilih 10 molekul yang paling berbeda rankingnya dan analisis kemungkinan penyebab berdasarkan fitur.
12. Buat pembahasan tentang mengapa screening komputasi harus diperlakukan sebagai prioritas, bukan bukti kinerja final.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Unduh dan pahami 350 record HOPV15.	data dictionary + clean table
2	Model baseline/RF dan split terkunci.	model_data + notebook
3	Ranking top-20, Spearman, error/rank analysis.	ranking.csv + 5+ grafik
4	Buku penapisan virtual material.	buku + ranking + kode

Gambar/tabel analisis minimum yang wajib muncul

- distribusi PCE
- HOMO-LUMO-gap scatter
- predicted vs observed PCE
- predicted rank vs experimental rank

Struktur minimum buku proyek

- Bab 1. Informatika material dan HOPV
- Bab 2. Data elektronik dan endpoint PCE
- Bab 3. Model dan metode ranking
- Bab 4. Hasil penapisan dan validasi
- Bab 5. Ketidakpastian dan keputusan material

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.

- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 9

ENUMERASI PUSTAKA MOLEKUL VIRTUAL: 125 STRUKTUR DARI 5 CORE × 5×5 VARIASI R-GROUP

Kaitan dengan materi kuliah

Bab 9: kimia kombinatorial, enumerasi, desain multiobjektif, Pareto dan seleksi keragaman.

Pertanyaan proyek yang harus dijawab

Bagaimana perubahan substituen sederhana mengubah ruang sifat suatu pustaka virtual, dan bagaimana memilih subset yang seimbang tanpa melakukan sintesis?

Objek/dataset yang langsung digunakan

Proyek ini sepenuhnya digital. Gunakan lima core dan lima pilihan substituen R pada dua posisi substitusi konseptual. Target teoritis $5 \times 5 \times 5 = 125$ struktur. Tidak ada prosedur sintesis, kondisi reaksi, atau penggunaan bahan fisik.

Kode	Core digital	SMILES dasar
Core A	benzene 1,4-disubstitusi	<chem>[*:1]c1ccc([*:2])cc1</chem>
Core B	furan disubstitusi	<chem>[*:1]c1cc([*:2])oc1</chem>
Core C	pyrrole disubstitusi	<chem>[*:1]c1cc([*:2])[nH]c1</chem>
Core D	thiophene disubstitusi	<chem>[*:1]c1cc([*:2])sc1</chem>
Core E	cyclohexane 1,4-disubstitusi	<chem>[*:1]C1CCC([*:2])CC1</chem>

Kode	Substituen digital	Fragment SMILES	Keterangan
R1	H	<chem>[H]</chem>	hidrogen
R2	OH	<chem>O</chem>	hidroksi
R3	CH3	<chem>C</chem>	metil
R4	OCH3	<chem>OC</chem>	metoksi
R5	NH2	<chem>N</chem>	amino

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
RDKit	Buka sumber resmi	manipulasi graf, validasi, descriptor dan fingerprint
PubChem	Buka sumber resmi	opsional: cek apakah sebagian struktur sudah dikenal

Cara memperoleh data dan memulai pada hari pertama

- Salin lima core dengan atom dummy [*:1] dan [*:2] serta lima fragment substituen dari tabel ke file templates.csv.

- Uji pemasangan substituen pada Core A sampai menghasilkan 25 struktur unik/valid; setelah aturan attachment benar, terapkan pada empat core lainnya.
- Semua struktur hanya objek digital; proyek tidak memuat kondisi sintesis, pereaksi laboratorium, atau prosedur pembuatan fisik.

Berkas yang harus dibuat kelompok

- library_125_raw.smi
- library_125_clean.csv
- library_descriptors.csv
- pareto_subset.csv
- notebook_p9.ipynb

Kolom/variabel minimum

- library_id
- core
- R1
- R2
- SMILES
- valid
- duplicate
- MolWt
- LogP
- TPSA
- HBD
- HBA
- rotatable bonds
- fingerprint
- Pareto flag

Prosedur rinci

1. Tetapkan aturan graf untuk dua titik substitusi pada masing-masing core; dokumentasikan atom indeks/titik attachment.
2. Enumerasikan seluruh kombinasi 5×5 substituen untuk tiap core sehingga terbentuk maksimal 125 struktur.
3. Simpan semua hasil sebelum pembersihan, termasuk struktur gagal.
4. Parse RDKit dan hapus hanya struktur yang tidak valid setelah dicatat dalam log.
5. Canonicalize SMILES dan lakukan deduplikasi.
6. Hitung descriptor MolWt, LogP, TPSA, HBD, HBA, rotatable bonds dan ring count.

7. Buat visualisasi ruang descriptor dan fingerprint similarity.
8. Definisikan tujuan multiobjektif yang netral untuk pembelajaran, misalnya meminimalkan MW ekstrem dan LogP ekstrem sambil menjaga keragaman struktur; tuliskan rumusan sebelum melihat hasil.
9. Bangun Pareto front atau skor multi-kriteria yang transparan.
10. Pilih 20 struktur representatif yang mencakup semua core dan menunjukkan keragaman descriptor.
11. Gambar grid 20 struktur pilihan dan tabel alasan pemilihan.
12. Lakukan pencarian exact/similarity PubChem hanya untuk mengetahui novelty basis data, bukan untuk mengklaim struktur baru secara ilmiah.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Definisikan core/substituen dan buat enumerator untuk satu core.	pseudocode + 25 struktur uji
2	Enumerasi 125, validasi dan deduplikasi.	library clean + log gagal/duplikat
3	Descriptor, Pareto, diversity dan 20 struktur pilihan.	pareto subset + visualisasi
4	Buku desain pustaka virtual.	buku + library + kode

Gambar/tabel analisis minimum yang wajib muncul

- diagram combinatorial enumeration
- PCA/descriptor map 125 struktur
- Pareto plot
- grid 20 struktur terpilih

Struktur minimum buku proyek

- Bab 1. Enumerasi dan ruang kombinatorial
- Bab 2. Aturan core dan R-group
- Bab 3. Validasi 125 struktur
- Bab 4. Deskriptor, Pareto dan keragaman
- Bab 5. Keterbatasan dan etika klaim novelty

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.

- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 10

LABORATORIUM KIMIA DIGITAL: ELN/LIMS UNTUK 30 JOB KOMPUTASI MOLEKUL

Kaitan dengan materi kuliah

Bab 9–10: ELN, LIMS, audit trail, otomasi, reproducibility dan siklus Design–Make–Test–Analyze dalam konteks komputasi.

Pertanyaan proyek yang harus dijawab

Bagaimana merancang sistem informasi laboratorium yang dapat melacak sampel, job komputasi, file hasil, versi metode dan keputusan untuk 30 molekul?

Objek/dataset yang langsung digunakan

Gunakan 30 molekul yang sama dengan Proyek 1. Buat Sample ID IK-001 sampai IK-030 sesuai urutan tabel Proyek 1. Setiap sample menjalani empat job digital: J1 verifikasi identifier, J2 gambar 2D + descriptor, J3 generasi satu struktur 3D, J4 quality-control dan finalisasi record.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
PubChem	Buka sumber resmi	data sumber 30 molekul
RDKit	Buka sumber resmi	job komputasi
SQLite	Buka sumber resmi	basis data ELN/LIMS sederhana

Cara memperoleh data dan memulai pada hari pertama

- Impor 30 molekul Proyek 1 ke sample_master.csv dan tetapkan IK-001 s.d. IK-030 secara permanen.
- Buat database kosong dan enam tabel sebelum menjalankan job; kemudian isi hanya lima sample pertama sebagai pilot.
- Setelah status/audit trail berfungsi, jalankan seluruh 30 sample dan ekspor laporan dashboard.

Berkas yang harus dibuat kelompok

- lims.sqlite
- sample_master.csv
- job_log.csv
- audit_log.csv
- SOP_ELN_LIMS.pdf
- notebook_p10.ipynb

Kolom/variabel minimum

- sample_id
- molecule_name
- source_cid
- job_id
- job_type
- operator/member
- start_time
- end_time
- software_version
- input_file
- output_file
- status
- qc_status
- decision
- audit_timestamp

Prosedur rinci

1. Buat master 30 sample ID IK-001...IK-030 dan hubungkan dengan nama molekul.
2. Rancang tabel SQLite minimal: sample, job, result, file_registry, audit_log, member.
3. Tetapkan status job: queued, running, completed, failed, reviewed.
4. Tuliskan SOP digital untuk empat tipe job J1–J4.
5. Jalankan J1 dan J2 untuk seluruh 30 molekul; catat waktu, versi perangkat dan nama anggota.
6. Jalankan J3 untuk minimal 15 molekul dan catat output SDF/struktur 3D.
7. Jalankan J4 sebagai pemeriksaan lintas anggota: reviewer tidak boleh sama dengan operator job utama untuk minimal 10 sample.
8. Setiap perubahan record penting harus menghasilkan audit_log, misalnya perubahan SMILES atau status QC.

9. Buat dashboard sederhana dari SQL/pandas: jumlah job per status, waktu rata-rata, jumlah failure, jumlah record menunggu review.

10. Simulasikan satu kegagalan komputasi yang tidak berbahaya, misalnya parse SMILES yang sengaja dibuat tidak lengkap pada COPY kerja, lalu dokumentasikan penanganannya; data sumber asli tidak boleh diubah.

11. Buat prosedur backup/versi dan struktur folder.

12. Lakukan audit oleh kelompok lain terhadap lima sample acak.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Desain skema, sample master dan SOP J1–J4.	ERD + sample_master + SOP draft
2	Isi J1/J2 seluruh 30 sample, J3 minimal 15.	database + job log + files
3	QC silang, audit log, dashboard dan audit 5 sample.	audit report + dashboard
4	Buku “Laboratorium Kimia Digital”.	buku + SQLite + SOP + presentasi

Gambar/tabel analisis minimum yang wajib muncul

- ERD LIMS
- state diagram status job
- dashboard job status
- contoh audit trail satu sample

Struktur minimum buku proyek

- Bab 1. Konsep ELN/LIMS
- Bab 2. Desain sample dan job
- Bab 3. Implementasi SQLite dan SOP
- Bab 4. Audit trail, QC dan dashboard
- Bab 5. Reproducibility dan human oversight

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 11

INFORMATIKA REAKSI BIOKIMIA: 12 REAKSI GLIKOLISIS/TCA DENGAN RHEA DAN ChEBI

Kaitan dengan materi kuliah

Bab 2 dan 9: representasi reaksi, atom/participant mapping konseptual, basis data reaksi, interoperabilitas dan retrosintesis/transformasi secara komputasional.

Pertanyaan proyek yang harus dijawab

Bagaimana reaksi biokimia direpresentasikan sebagai objek data yang memiliki participant, identifier, arah dan keterhubungan jaringan?

Objek/dataset yang langsung digunakan

No.	Reaksi tetap	Persamaan ringkas
1	Glucose phosphorylation	D-glucose + ATP → D-glucose 6-phosphate + ADP + H ⁺
2	Phosphoglucose isomerization	D-glucose 6-phosphate ↔ D-fructose 6-phosphate
3	Phosphofructokinase step	D-fructose 6-phosphate + ATP → fructose 1,6-bisphosphate + ADP + H ⁺
4	Fructose-bisphosphate cleavage	fructose 1,6-bisphosphate ↔ glyceraldehyde 3-phosphate + dihydroxyacetone phosphate
5	Triose-phosphate isomerization	dihydroxyacetone phosphate ↔ glyceraldehyde 3-phosphate
6	GAP oxidation/phosphorylation	glyceraldehyde 3-phosphate + NAD ⁺ + phosphate ↔ 1,3-bisphosphoglycerate + NADH + H ⁺
7	Phosphoglycerate kinase	1,3-bisphosphoglycerate + ADP ↔ 3-phosphoglycerate + ATP
8	Phosphoglycerate mutase	3-phosphoglycerate ↔ 2-phosphoglycerate
9	Enolase	2-phosphoglycerate ↔ phosphoenolpyruvate + H ₂ O
10	Pyruvate kinase	phosphoenolpyruvate + ADP + H ⁺ → pyruvate + ATP
11	Lactate dehydrogenase	pyruvate + NADH + H ⁺ ↔ lactate + NAD ⁺
12	Citrate synthase	oxaloacetate + acetyl-CoA + H ₂ O → citrate + CoA

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
Rhea	Buka sumber resmi	Rhea ID, persamaan, arah reaksi, ChEBI participant dan EC link
Rhea download	Buka sumber resmi	bulk TSV/RDF/format reaksi
ChEBI	Buka sumber resmi	identifier kimia participant

Cara memperoleh data dan memulai pada hari pertama

- Buka Rhea dan cari satu reaksi pertama “D-glucose phosphorylation”; catat Rhea ID, arah, serta semua participant ChEBI.
- Gunakan format tabel yang sama untuk 11 reaksi berikutnya. Jangan mengarang Rhea ID: setiap ID harus diverifikasi pada halaman Rhea.
- Setelah 12 reaksi lengkap, baru bangun network reaction-metabolite.

Berkas yang harus dibuat kelompok

- rhea_12_reactions.csv
- participants.csv
- reaction_network.csv
- reaction_graph.graphml
- notebook_p11.ipynb

Kolom/variabel minimum

- reaction_name
- Rhea ID
- direction
- participant name
- ChEBI ID
- stoichiometric side
- EC number/link if available
- reaction connectivity

Prosedur rinci

1. Cari 12 reaksi pada Rhea berdasarkan nama/persamaan dan catat Rhea ID yang cocok; verifikasi persamaan manual.
2. Ekspor/catat seluruh participant dan ChEBI ID masing-masing reaksi.
3. Pisahkan data menjadi tabel reaction dan participant.
4. Normalisasi participant sehingga nama yang sama menggunakan ChEBI ID yang sama.
5. Buat daftar input-output untuk setiap reaksi tanpa memasukkan kondisi eksperimen atau prosedur laboratorium.

6. Buat network reaction–metabolite: node metabolit dan node reaksi, edge menunjukkan partisipasi.
7. Hitung degree sederhana dan identifikasi metabolit yang menjadi hub pada subset 12 reaksi.
8. Buat alur urutan 10 reaksi glikolisis dan tunjukkan keterhubungan dengan lactate/citrate branch yang dipilih.
9. Bandingkan bagaimana nama biasa, ChEBI ID dan struktur/SMILES (jika tersedia) berperan dalam interoperabilitas.
10. Pilih tiga reaksi dan buat diagram data: participant → identifier → reaction → pathway context.
11. Dokumentasikan ambiguity: proton, air, arah reaksi atau bentuk ion yang ditampilkan berbeda antar sumber.
12. Susun skema basis data kecil untuk menyimpan reaction network ini.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Temukan 12 Rhea ID dan participant/ChEBI ID.	rhea_12_reactions + verifikasi
2	Normalisasi reaction/participant dan buat database tabel.	participants + reaction table
3	Bangun network, analisis hub dan ambiguity.	GraphML + network figures
4	Buku informatika reaksi dan tutorial Rhea.	buku + data + graph

Gambar/tabel analisis minimum yang wajib muncul

- diagram 12 reaction network
- skema reaction-participant database
- contoh satu record Rhea→ChEBI
- hub metabolite plot

Struktur minimum buku proyek

- Bab 1. Reaksi sebagai objek informasi
- Bab 2. Rhea dan ChEBI
- Bab 3. Dataset 12 reaksi
- Bab 4. Jaringan reaksi dan interoperabilitas
- Bab 5. Ambiguitas dan batas representasi

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.

- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 12

AI GENERATIF YANG DAPAT DIAUDIT: SELFIES MUTATION PADA 5.000 MOLEKUL QM9

Kaitan dengan materi kuliah

Bab 7, 9 dan 10: representasi terpelajar/generatif, novelty, validitas, chemical space, bias dan human oversight.

Pertanyaan proyek yang harus dijawab

Apakah mutasi berbasis SELFIES dapat menghasilkan string molekul yang valid sekaligus tetap berada dekat dengan domain kimia data awal?

Objek/dataset yang langsung digunakan

Ambil 5.000 molekul QM9 dengan random_state=2026. Pilih 100 seed dari subset tersebut. Encode SMILES→SELFIES; buat tepat 10 mutasi string per seed sehingga dihasilkan sekitar 1.000 kandidat. Proyek berhenti pada evaluasi digital; tidak ada sintesis atau uji biologis.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
QM9	Buka sumber resmi	data latih/referensi
SELFIES docs	Buka sumber resmi	encoder/decoder dan tokenisasi
RDKit	Buka sumber resmi	parse, descriptor, fingerprint dan chemical-space comparison

Cara memperoleh data dan memulai pada hari pertama

- Gunakan 5.000 ID/SMILES QM9 yang disampling dengan random_state=2026 dan simpan file train_ids.txt.
- Instal package selfies; uji encode→decode pada 20 SMILES sebelum membuat mutasi.
- Gunakan random seed tetap 2026 untuk pemilihan seed dan mutasi agar 1.000 kandidat dapat direproduksi.

Berkas yang harus dibuat kelompok

- qm9_train_5000.csv
- seed_100.csv
- generated_1000.csv
- generation_metrics.csv
- notebook_p12.ipynb

Kolom/variabel minimum

- seed_id
- seed_smiles
- seed_selfies
- mutation_id
- generated_selfies
- decoded_smiles
- RDKit_valid
- unique
- novel_vs_train
- heavy_atoms
- MolWt
- LogP
- TPSA
- similarity_to_seed

Prosedur rinci

1. Ambil 5.000 QM9 record dan simpan daftar ID/SMILES.
2. Pilih 100 seed secara acak dengan random_state=2026.
3. Encode semua seed ke SELFIES.
4. Bangun prosedur mutasi token sederhana yang terdokumentasi, misalnya penggantian/penyisipan/penghapusan satu token sesuai fungsi library; simpan random seed.
5. Buat 10 kandidat per seed.
6. Decode SELFIES ke SMILES dan uji parse RDKit.
7. Hitung validity, uniqueness dan novelty terhadap 5.000 data awal.
8. Batasi analisis utama pada kandidat yang tetap menggunakan unsur CHONF dan jumlah heavy atoms ≤ 9 agar domain tetap sejalan dengan QM9.
9. Hitung descriptor dasar dan similarity ke seed.

10. Bandingkan distribusi descriptor seed/training/generated.
11. Buat PCA training vs generated dan tandai kandidat outlier.
12. Pilih 20 kandidat contoh: 10 dekat domain dan 10 outlier; jelaskan mengapa validitas sintaksis tidak otomatis berarti relevansi ilmiah.
13. Tulis deklarasi human oversight dan batas klaim: kandidat hanya objek komputasi.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Subset QM9, 100 seed dan pipeline SELFIES.	train + seed + encoder test
2	Generate 1.000 kandidat dan hitung validity/uniqueness/novelty.	generated CSV + metrics
3	Descriptor, PCA, outlier dan galeri 20 kandidat.	chemical-space figures + analysis
4	Buku AI generatif yang dapat diaudit.	buku + code + data

Gambar/tabel analisis minimum yang wajib muncul

- workflow SMILES→SELFIES→mutation→SMILES
- bar validity/uniqueness/novelty
- PCA train vs generated
- galeri 20 kandidat digital

Struktur minimum buku proyek

- Bab 1. Bahasa molekul untuk AI
- Bab 2. QM9 dan SELFIES
- Bab 3. Metode generasi yang dapat direproduksi
- Bab 4. Validity, novelty dan chemical space
- Bab 5. Bias, outlier dan human oversight

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.

- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 13

ATLAS PRODUK ALAM TERBUKA: 500 MOLEKUL COCONUT + 20 MOLEKUL REFERENSI

Kaitan dengan materi kuliah

Bab 2, 4, 5 dan 6: data produk alam, standardisasi, descriptor, fingerprint, chemical space, similarity dan scaffold.

Pertanyaan proyek yang harus dijawab

Bagaimana keragaman struktur produk alam terlihat melalui descriptor, scaffold dan fingerprint, dan di mana posisi 20 molekul referensi yang dikenal?

Objek/dataset yang langsung digunakan

Unduh COCONUT dan pilih 500 record dengan aturan: valid RDKit, $100 \leq MW \leq 700$, no metal, canonical SMILES tersedia; sample random_state=2026. Selain itu, buat daftar referensi 20 molekul berikut dan cocokkan ke PubChem/COCONUT jika tersedia: Vanilin, Eugenol, Limonena, Mentol, Timol, Sinamaldehyda, Linalool, Geraniol, Kuersetin, Kaempferol, Katekin, Asam galat, Asam kafeat, Asam ferulat, Kurkumin, Resveratrol, Rutin, Naringenin, Hesperetin, Asam rosmarinat.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
COCONUT Download	Buka sumber resmi	500 produk alam
PubChem	Buka sumber resmi	verifikasi 20 molekul referensi
RDKit	Buka sumber resmi	descriptor, scaffold dan fingerprint

Cara memperoleh data dan memulai pada hari pertama

- Unduh COCONUT CSV, lakukan filter yang tertulis, lalu sample 500 dengan random_state=2026 dan simpan ID.
- Buat reference_20.csv berisi 20 nama yang telah ditetapkan; cari PubChem CID dan, bila ada, COCONUT ID untuk masing-masing.
- Peta PCA dan nearest-neighbor baru dibuat setelah 500+20 record lolos QC struktur.

Berkas yang harus dibuat kelompok

- coconut_500.csv
- reference_20.csv
- np_descriptors.csv

- np_clusters.csv
- atlas_structures.pdf
- notebook_p13.ipynb

Kolom/variabel minimum

- COCONUT ID
- name
- SMILES
- MW
- LogP
- TPSA
- HBD
- HBA
- ring count
- fraction Csp3
- Murcko scaffold
- cluster
- reference flag

Prosedur rinci

1. Ambil dan dokumentasikan 500 record COCONUT.
2. Buat 20 reference record dari daftar yang ditetapkan; hubungkan ke PubChem CID dan COCONUT ID bila ada.
3. Canonicalize struktur dan hitung descriptor.
4. Hitung Morgan fingerprint untuk seluruh 520 record atau gabungan tanpa duplikasi.
5. Hitung Murcko scaffold dan frekuensinya.
6. Lakukan PCA descriptor dan clustering sederhana.
7. Plot 20 reference di atas peta PCA 500 produk alam.
8. Untuk setiap reference, cari 3 tetangga terdekat berdasarkan Tanimoto pada subset 500.
9. Buat tabel top-10 scaffold yang paling sering.
10. Buat atlas struktur untuk 20 reference + 30 struktur representatif cluster/scaffold.
11. Bandingkan apakah kedekatan descriptor selalu sejalan dengan kedekatan fingerprint.
12. Diskusikan bias dataset agregat: sumber, organism metadata, duplikasi sejarah dan kelengkapan nama.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Subset 500 + verifikasi 20 reference.	coconut_500 + reference_20
2	Descriptor, fingerprint, scaffold.	feature tables
3	PCA, clustering, nearest neighbors, atlas.	figures + atlas + clusters
4	Buku Atlas Informatika Produk Alam.	buku + data + kode

Gambar/tabel analisis minimum yang wajib muncul

- PCA 500 + 20 reference
- top scaffold bar chart
- nearest-neighbor panels
- atlas 50 struktur

Struktur minimum buku proyek

- Bab 1. Produk alam sebagai data
- Bab 2. COCONUT dan kurasi
- Bab 3. Descriptor, fingerprint dan scaffold
- Bab 4. Atlas dan tetangga terdekat
- Bab 5. Bias dataset dan interpretasi

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 14

INFORMATIKA KIMIA LINGKUNGAN MOLEKUL-AIR: QSPR ENERGI HIDRASI FREESOLV

Kaitan dengan materi kuliah

Bab 5–7: deskriptor, hubungan struktur-sifat, regresi, validasi, ketidakpastian dan domain aplikasi.

Pertanyaan proyek yang harus dijawab

Fitur struktur apa yang paling berhubungan dengan energi bebas hidrasi, dan seberapa baik model QSPR sederhana memprediksi nilai eksperimen?

Objek/dataset yang langsung digunakan

Gunakan dataset FreeSolv versi MoleculeNet/DeepChem atau repository resmi. Target minimal seluruh record yang mempunyai SMILES dan experimental hydration free energy; pada dokumentasi DeepChem saat ini, dataset tersebut memuat sekitar 643 molekul. Gunakan endpoint experimental hydration free energy dalam satuan yang disediakan dataset dan jangan mengubah unit tanpa dokumentasi. Jika jumlah record berbeda karena versi sumber berubah, catat versi dan tanggal akses.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
FreeSolv official repository	Buka sumber resmi	data eksperimen + struktur
DeepChem MoleculeNet	Buka sumber resmi	opsi loader FreeSolv
RDKit	Buka sumber resmi	descriptor

Cara memperoleh data dan memulai pada hari pertama

- Cara termudah: gunakan DeepChem `dc.molnet.load_freesolv(...)` atau file CSV dari repository FreeSolv.
- Pastikan kolom endpoint experimental hydration free energy dikenali dan satuannya dicatat dari sumber; jangan mencampur nilai eksperimen dan hasil kalkulasi.
- Buat satu notebook EDA pada 100 record, kemudian jalankan seluruh dataset setelah pipeline benar.

Berkas yang harus dibuat kelompok

- `freesolv_raw.csv`
- `freesolv_features.csv`
- `freesolv_results.csv`
- `freesolv_errors.csv`
- `notebook_p14.ipynb`

Kolom/variabel minimum

- compound ID
- SMILES
- experimental hydration free energy
- uncertainty if available
- MolWt
- LogP
- TPSA
- HBD
- HBA
- fraction Csp3
- ring count
- prediction
- residual

Prosedur rinci

1. Unduh/muat FreeSolv dan catat versi/sumber file.
2. Validasi SMILES dan endpoint numerik.
3. Hitung descriptor dasar.
4. Lakukan eksplorasi korelasi descriptor dengan hydration free energy.
5. Buat split 80/20 dengan random_state=2026 dan simpan assignment; bila mampu, bandingkan dengan scaffold split.
6. Bangun baseline mean predictor.
7. Bangun Linear Regression dan Random Forest.
8. Evaluasi MAE, RMSE dan R^2 .
9. Buat residual plot dan identifikasi 20 error terbesar.
10. Bandingkan error dengan MW, TPSA dan LogP untuk mencari pola failure.
11. Jika uncertainty eksperimen tersedia, tampilkan bagaimana error model dibandingkan dengan ketidakpastian pengukuran.
12. Susun pembahasan kimia tentang polaritas/hydrogen bonding dan interaksi molekul-air tanpa membuat klaim yang melampaui dataset.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	FreeSolv QA + descriptor.	raw + feature CSV
2	EDA, baseline, LR/RF.	model notebook + result table
3	Error analysis, uncertainty, random vs scaffold split bila mampu.	errors + figures
4	Buku QSPR hidrasi.	buku + data + kode

Gambar/tabel analisis minimum yang wajib muncul

- correlation heatmap
- predicted vs observed
- residual vs LogP/TPSA
- galeri error terbesar

Struktur minimum buku proyek

- Bab 1. Energi hidrasi dan data FreeSolv
- Bab 2. Descriptor molekul-air
- Bab 3. QSPR dan validasi
- Bab 4. Error dan ketidakpastian
- Bab 5. Domain aplikasi dan keterbatasan

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.
- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

PROYEK 15

OPEN CHEMOINFORMATICS REPRODUCIBLE PACKAGE: REPLIKASI TERKUNCI MODEL ESOL OLEH KELOMPOK LAIN

Kaitan dengan materi kuliah

Bab 1–10, khususnya open science, version control, reproducibility, environment, seed, checksum, model card dan audit lintas kelompok.

Pertanyaan proyek yang harus dijawab

Dapatkah kelompok lain menjalankan pipeline yang sama dan memperoleh data split, metrik, gambar dan kesimpulan yang sama tanpa bertanya kepada pembuat?

Objek/dataset yang langsung digunakan

Gunakan ESOL 1.128 record sebagai benchmark yang sama untuk semua kelompok yang mengambil proyek ini. Pipeline wajib dikunci: canonicalization RDKit; descriptor MolWt, MolLogP, TPSA, HBD, HBA, RotB, RingCount, FractionCSP3; model RandomForestRegressor dengan `n_estimators=300` dan `random_state=2026`; split train/test menggunakan assignment CSV yang dibuat sekali dan diarsipkan. Tujuan proyek bukan mengejar skor terbaik, tetapi paket reproduksi.

Sumber data resmi

Sumber resmi	Alamat	Dipakai untuk
ESOL CSV	Buka sumber resmi	dataset tetap
RDKit	Buka sumber resmi	descriptor
scikit-learn	Buka sumber resmi	Random Forest dan metrik

Cara memperoleh data dan memulai pada hari pertama

- Unduh ESOL sekali dan simpan sebagai `data/raw/delaney-processed.csv`; hitung SHA-256 sebelum analisis.
- Buat `pipeline.py` sebelum notebook eksplorasi diselesaikan: tujuan proyek adalah reproduksi otomatis, bukan hanya notebook interaktif.
- Tentukan satu kelompok pasangan sebagai reproducer sejak Pekan 1 agar uji lintas kelompok tidak terlambat.

Berkas yang harus dibuat kelompok

- `README.md`
- `environment.yml` atau `requirements.txt`
- `data/raw/delaney-processed.csv`
- `data/processed/esol_features.csv`
- `splits/split_assignment.csv`

- src/pipeline.py
- notebooks/analysis.ipynb
- results/metrics.json
- figures/pred_vs_obs.png
- checksums.sha256
- MODEL_CARD.md
- REPRODUCTION_REPORT.md

Kolom/variabel minimum

- file path
- version
- checksum
- seed
- parameter
- expected output
- observed output by reproducer
- match/mismatch
- resolution

Prosedur rinci

1. Buat struktur folder standar sebelum analisis dimulai.
2. Simpan file ESOL mentah dan checksum-nya; file raw tidak boleh diubah.
3. Tulis satu script pipeline.py yang dapat menjalankan proses dari raw → features → split → model → metrics → figures.
4. Kunci semua parameter dan random_state=2026.
5. Buat split_assignment.csv dan jangan membuat ulang split secara acak di tiap run.
6. Simpan metrics.json dan file gambar dengan nama tetap.
7. Tulis README mulai dari instalasi environment sampai satu perintah utama menjalankan pipeline.
8. Simpan versi Python dan package utama.
9. Buat checksum untuk file penting.
10. Serah-terimakan paket kepada kelompok pasangan; kelompok pasangan menjalankan tanpa bantuan lisan selama percobaan pertama.
11. Kelompok reproducer mengisi REPRODUCTION_REPORT: tahap berhasil, gagal, perbedaan angka, waktu, dan penyebab.

12. Perbaiki dokumentasi/kode berdasarkan laporan tersebut, lalu lakukan run kedua.
13. Bandingkan hash/angka output yang seharusnya deterministik dan jelaskan output yang mungkin berbeda karena versi platform.
14. Susun model card dan data card final.

Rencana kerja 4 pekan

Pekan	Pekerjaan konkret	Bukti yang harus diserahkan
1	Struktur folder, raw checksum, features, split terkunci.	repo/package draft + checksum
2	pipeline.py deterministik + expected outputs.	metrics + figures + README v1
3	Reproduksi oleh kelompok pasangan dan perbaiki.	REPRODUCTION_REPORT v1/v2
4	Buku tentang reproduksibilitas + paket final.	buku + package yang dapat dijalankan ulang

Gambar/tabel analisis minimum yang wajib muncul

- dependency/workflow diagram
- run manifest
- expected vs reproduced metrics
- timeline failure→fix

Struktur minimum buku proyek

- Bab 1. Reproducible research
- Bab 2. Dataset dan pipeline terkunci
- Bab 3. Environment, seed dan checksum
- Bab 4. Uji reproduksi lintas kelompok
- Bab 5. Pelajaran, model card dan data card

Pembagian kerja lima anggota yang disarankan

- Anggota 1: koordinator, rumusan masalah, validasi konsep dan integrasi naskah.
- Anggota 2: akuisisi data, provenance, data dictionary dan kurasi manual.
- Anggota 3: kode/pipeline RDKit-Python/SQL sesuai proyek dan README.
- Anggota 4: analisis statistik, visualisasi, validasi serta error analysis.
- Anggota 5: atlas/gambar, tabel, format buku, glosarium, sitasi dan presentasi.

Kriteria proyek dianggap selesai

- Seluruh dataset/objek yang ditentukan di atas benar-benar diproses, bukan sekadar dibahas secara teoritis.
- Notebook/script dapat dijalankan kembali dari data mentah atau input yang terdokumentasi.
- Semua gambar dan tabel dalam buku dapat ditelusuri ke file hasil analisis.
- Ada minimal satu bagian validasi atau error/failure analysis, bukan hanya hasil yang berhasil.

- Kesimpulan tidak boleh melebihi bukti yang tersedia pada dataset.

CATATAN DOSEN: PEMILIHAN PROYEK DAN KESERAGAMAN BEBAN

Untuk menjaga beban antar-kelompok tetap setara, satu kelompok hanya mengambil satu proyek. Proyek 1–5 cocok untuk mahasiswa yang baru menguasai Python/RDKit; Proyek 6–8 dan 14 menuntut analisis statistik/model; Proyek 9, 11–12 menuntut manipulasi representasi/graph/string; Proyek 10 dan 15 berfokus pada rekayasa data dan reproduksibilitas. Dosen dapat membatasi satu judul untuk satu kelompok agar produk akhir menjadi 15 buku yang saling melengkapi.

F. STRUKTUR WAJIB BUKU SETIAP KELOMPOK

Setiap buku menggunakan struktur berikut.

Bagian awal

- Cover
- Halaman judul
- Identitas anggota kelompok
- Kata Pengantar
- Capaian Pembelajaran
- Daftar Isi
- Daftar Gambar
- Daftar Tabel

Bagian isi

Bab 1 — Landasan Teori dan Permasalahan

Menjelaskan latar belakang, tujuan, konsep kimia dan informatika kimia.

Bab 2 — Data, Representasi dan Metodologi

Menjelaskan sumber data, lisensi, identifier, preprocessing dan metode.

Bab 3 — Implementasi Proyek

Menjelaskan prosedur, algoritme, kode dan parameter.

Bab 4 — Hasil dan Pembahasan

Memuat tabel, grafik, struktur, visualisasi dan interpretasi kimia.

Bab 5 — Validasi, Keterbatasan dan Kesimpulan

Membahas reliabilitas, ketidakpastian, keterbatasan dan rekomendasi.

Jumlah bab boleh lebih dari lima.

Bagian pembelajaran

- Praktikum/tutorial
- Contoh soal
- minimal 10 soal latihan
- rangkuman
- glosarium

Bagian akhir

- Kesimpulan
 - Daftar pustaka
 - Lampiran dataset
 - Lampiran kode
 - Data card
 - Pembagian tugas
 - Log aktivitas
 - Pernyataan penggunaan AI
-

G. DATA CARD WAJIB

Setiap kelompok harus membuat **Data Card** yang menjelaskan:

Komponen	Informasi
Nama dataset	...
Sumber	...
URL/DOI	...
Tanggal akses	...
Lisensi	...
Jumlah awal	...
Jumlah akhir	...
Jenis molekul	...
Identifier	...
Endpoint	...
Unit	...
Missing value	...
Duplikasi	...
Aturan kurasi	...

Komponen	Informasi
Keterbatasan	...

H. CHECKLIST QUALITY CONTROL

Sebelum buku dikumpulkan, kelompok wajib memastikan:

Struktur molekul

- semua SMILES dapat dibaca perangkat;
- valensi diperiksa;
- muatan diperiksa;
- aromatisitas diperiksa;
- stereokimia diperiksa bila relevan.

Identitas

- duplikasi sudah ditangani;
- garam/solvat dijelaskan;
- tautomer diperlakukan konsisten;
- parent compound didefinisikan.

Data

- sumber jelas;
- lisensi jelas;
- tanggal akses tercatat;
- unit konsisten;
- perubahan data terdokumentasi.

Model

Jika proyek menggunakan machine learning:

- train/test tidak mengalami leakage;
- baseline tersedia;
- seed dicatat;
- preprocessing hanya di-*fit* pada data latih;
- metrik sesuai tujuan;
- domain aplikasi dibahas.

Reproduksibilitas

- data tersedia;

- kode tersedia;
- versi software dicatat;
- README tersedia;
- parameter tersedia;
- hasil dapat diperiksa.

I. RUBRIK PENILAIAN

Aspek	Bobot
Ketepatan konsep kimia	20%
Kualitas data dan provenance	15%
Metode dan validasi	20%
Reprodusibilitas	15%
Analisis dan argumentasi	15%
Komunikasi ilmiah dan kualitas buku	10%
Kolaborasi dan integritas	5%
TOTAL	100%

Interpretasi

Ketepatan konsep kimia: struktur, istilah, mekanisme, satuan dan interpretasi harus benar.

Data dan provenance: sumber, lisensi, metadata dan perubahan data harus dapat ditelusuri.

Metode dan validasi: metode harus sesuai pertanyaan penelitian dan dilengkapi validasi.

Reprodusibilitas: kode, environment, parameter dan petunjuk kerja memungkinkan pengulangan.

Analisis: kesimpulan harus berasal dari data dan tidak melebihi bukti.

Komunikasi: buku harus terstruktur, gambar dan tabel jelas, serta sitasi konsisten.

Kolaborasi: seluruh anggota memberikan kontribusi dan mencatat pekerjaannya.

J. PENILAIAN PROSES EMPAT PEKAN

Untuk mencegah buku dibuat hanya pada hari-hari terakhir, nilai proyek dapat dibagi:

Komponen	Bobot
Proposal dan outline — Pekan 1	10%

Komponen	Bobot
Dataset, kurasi dan metode — Pekan 2	20%
Hasil, analisis dan draft buku — Pekan 3	25%
Buku final — Pekan 4	30%
Presentasi dan tanya jawab	10%
Peer assessment kontribusi anggota	5%

K. MEKANISME PRESENTASI

Setiap kelompok mendapat waktu sekitar **15 menit**:

- 2 menit: masalah dan tujuan;
- 3 menit: data;
- 3 menit: metode;
- 4 menit: hasil terpenting;
- 2 menit: keterbatasan;
- 1 menit: kesimpulan.

Dilanjutkan tanya jawab.

Dosen dapat memilih satu anggota secara acak untuk menjelaskan kode, data, metode atau isi buku sehingga setiap mahasiswa harus memahami keseluruhan proyek.

L. ATURAN PENGGUNAAN AI

AI generatif diperbolehkan sebagai **alat bantu**, tetapi setiap kelompok wajib mencantumkan:

1. AI yang digunakan;
2. bagian pekerjaan yang dibantu AI;
3. contoh penggunaan utama;
4. cara melakukan verifikasi;
5. bagian yang diperbaiki manusia.

AI **tidak boleh** digunakan untuk:

- membuat data eksperimen palsu;
- menciptakan referensi fiktif;
- mengarang hasil komputasi;

- mengganti nilai hasil perhitungan;
- membuat klaim ilmiah yang tidak diverifikasi.

Setiap angka, referensi dan kesimpulan tetap menjadi tanggung jawab mahasiswa.

M. PRINSIP UTAMA YANG HARUS DINILAI

Proyek Informatika Kimia bukan lomba menghasilkan algoritme yang paling rumit.

Yang harus dapat dijawab mahasiswa adalah:

Mengapa data ini dipilih?

Mengapa representasi ini digunakan?

Mengapa metode ini sesuai?

Bagaimana hasil divalidasi?

Apa sumber ketidakpastiannya?

Dalam kondisi apa kesimpulan dapat gagal?

Dapatkan kelompok lain mengulang pekerjaan tersebut?

Dengan demikian produk akhirnya bukan sekadar kumpulan tulisan, melainkan **buku ilmiah mahasiswa yang merupakan dokumentasi lengkap sebuah proyek informatika kimia berbasis data, komputasi, validasi, interpretasi kimia, dan reproduksibilitas.**