A Precedent-Guided Co-Scientist for Side-Effect-Aware Drug Redesign

Yujin Kim [1] **Charmgil Hong** [1]

Abstract

We propose PRECEDE, a precedent-guided co-scientist for side-effect-aware drug redesign that revises a parent compound to mitigate a specified side effect while preserving therapeutic function. Rather than isolated molecular generation, PRECEDE frames redesign as evidence-grounded reasoning over drug–side-effect associations, biomedical knowledge graphs, and precedents of safety-driven optimization, coordinated by an LLM orchestrator with explicit policies and human-review checkpoints. We position PRECEDE as a human-supervised AI-for-science workflow in which hypotheses remain auditable, falsifiable, and bounded by prior pharmacology.

1. Motivation and Task Definition

Many drug candidates exhibit therapeutic activity yet carry side effects that require mitigation or restrict clinical use. Existing molecular optimization methods mainly target generic objectives, such as potency, drug-likeness, or broad ADMET properties (Jin et al., 2018; You et al., 2018). They typically treat each optimization instance as an independent generation problem. Pharmacological redesign, in contrast, often follows recognizable precedents, as in the transition from *tenofovir disoproxil fumarate* to *tenofovir alafenamide* (Ray et al., 2016). In these cases, safety liabilities are mitigated through interpretable structural or pharmacokinetic strategies, such as prodrug modification, bioisosteric replacement, and exposure modulation (Rautio et al., 2018; Kim et al., 2024). Recent agentic drug-discovery systems demonstrate the potential of LLM-based orchestration, retrieval, and tool use for molecular design (Averly et al., 2025; Liu et al., 2024). These systems search chemical space toward a target and a property profile. PRECEDE instead takes an approved drug and one documented side effect as input, classifies the effect by mechanism before any edit, and transfers a strategy abstracted from redesign precedents.

[1]Department of Computer Science and Electrical Engineering, Handong Global University, Pohang, Republic of Korea. Correspondence to: Charmgil Hong <charmgil@handong.edu>.

Accepted at the ICML 2026 Workshop on AI for Science (Poster).

We define *side-effect-aware drug redesign* as follows. Given a parent drug x, a target side effect a, and an optional therapeutic context c, the system produces ranked candidates $\{x'_i\}$ along with evidence trails, predicted efficacy–safety profiles, modification rationales, and an auditable redesign report. Unlike generic molecular generation, the task requires mechanism-informed screening of the side effect before any structural edit: the system must corroborate the drug–side-effect association, identify plausible liability factors, retrieve analogous precedents, and check whether candidate edits preserve the therapeutic mechanism.

PRECEDE addresses this task as a precedent-guided agentic workflow rather than a single generative model. It corroborates drug–side-effect associations using SIDER (Kuhn et al., 2016) and knowledge graphs such as PrimeKG (Chandak et al., 2023), classifies the side effect by likely mechanism, retrieves analogous precedents, and evaluates candidate edits with *in silico* efficacy and safety proxies. The hypothesis stage, from evidence grounding through *in silico* evaluation and reporting, is handled by the system, while experimental execution and final interpretation remain with human collaborators. This division keeps outputs as auditable hypotheses rather than autonomous recommendations.

This proposal contributes (i) a precedent memory that encodes historical redesign cases as structured records rather than isolated molecules, (ii) an attribution-aware decision policy that gates structural editing on the redesignability of the side effect, and (iii) a historical replay protocol that tests whether the reasoning of PRECEDE aligns with documented optimization decisions.

2. Proposed Approach

PRECEDE uses an LLM-based orchestrator to coordinate evidence retrieval, attribution, precedent search, constrained redesign, and candidate evaluation. Given a parent compound, a target side effect, and an optional therapeutic context, the orchestrator first checks whether external evidence supports the queried drug–side-effect association.

Attribution-aware routing. Before redesign, PRECEDE classifies the side effect into one of five categories based on established mechanism-based taxonomies (Edwards & Aronson, 2000; Aronson & Ferner, 2003; Park et al., 2011):

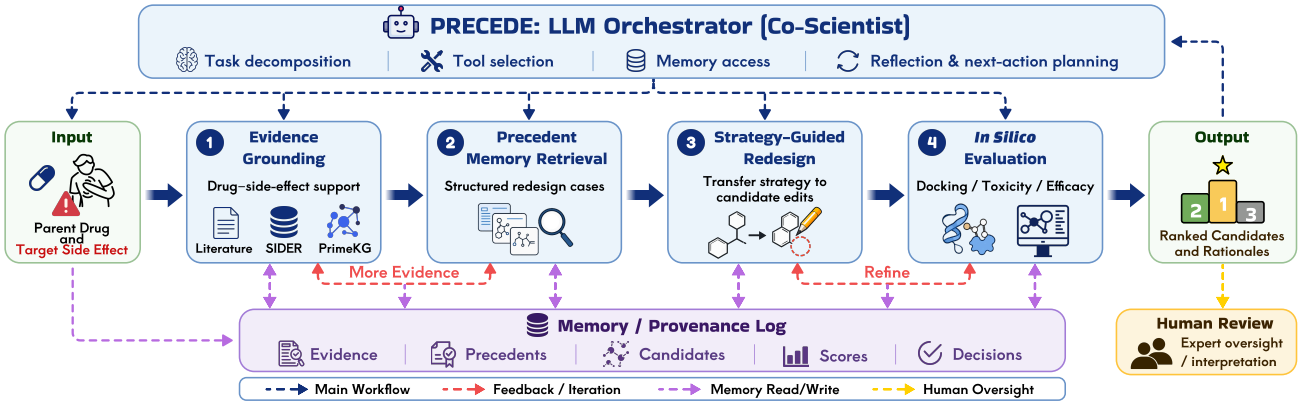

Figure 1. Overview of PRECEDE, an LLM-orchestrated workflow for evidence-grounded, precedent-guided drug redesign.

(i) target-mediated, (ii) off-target structural liability, (iii) metabolism or reactive intermediate, (iv) exposure or pharmacokinetic, and (v) insufficient evidence. Only categories (ii)–(iv) proceed to redesign; categories (i) and (v) are routed to human review. This routing keeps structural editing defensible and constrained to redesignable liabilities.

Precedent memory and strategy abstraction. Each precedent record encodes a parent compound, safety issue, structural modification, improved compound, and observed outcome. PRECEDE abstracts transferable strategies from these records, such as exposure modulation, liability-group replacement, pharmacophore preservation, and peripheral editing. The orchestrator selects strategies with precedents matching the attribution category and therapeutic context.

Constrained redesign and evaluation. Strategies guide a hybrid generation stack: matched molecular pair transformations from precedent memory, template-based bioisosteric replacements (Kim et al., 2024), and LLM-proposed substitutions. Candidate edits are filtered by chemical validity and synthetic accessibility. Surviving candidates are evaluated through target-specific docking, pharmacophore retention, toxicity predictors, and parent-similarity constraints.

Decision policy. Policies govern stage transitions: evidence below a confidence threshold (*e.g.*, no SIDER-confirmed association or fewer than two citations) triggers retrieval expansion; conflicting precedents are resolved by ranked alternatives scored on contextual similarity; failed evaluation (*e.g.*, docking degradation beyond a parent-relative margin or low synthetic accessibility) returns the system to redesign under tighter constraints. A provenance log records every evidence item, precedent, edit, score, and decision rationale; at termination, it exports an auditable redesign report.

3. Evaluation Roadmap

PRECEDE is evaluated along four axes that follow its workflow. *Evidence grounding* is assessed by support recall@k against SIDER and PrimeKG, by citation validity, and by attribution classification accuracy on a labeled subset of cases. *Precedent retrieval* is measured by top-k accuracy and strategy-level F1 on held-out cases. *Redesign quality* combines chemical validity, parent similarity, synthetic accessibility, predicted toxicity reduction, and therapeutic preservation via docking and pharmacophore retention as proxies where assay data are unavailable. *Expert review* provides a multi-rater check on rationale plausibility.

The main benchmark is historical replay on a pilot of 30–50 trajectories from prodrug case studies (Rautio et al., 2018) and matched molecular pair literature, each pairing a parent with its safety-driven successor. Holding out the successor makes each hypothesis falsifiable against a documented outcome: PRECEDE must recover the optimization from the parent and target side effect alone. We treat replay as necessary but not sufficient: agreement supports medicinal-chemistry reasoning, while interpretable disagreement remains informative. A proof-of-concept implementation and single-case walkthrough are in the appendix.

4. Governance and Expected Impact

PRECEDE is positioned as a co-scientist: a human-supervised hypothesis-generation system rather than an autonomous clinical or synthesis engine. It proposes evidence-grounded redesign hypotheses, while humans retain decision authority and scientific credit. It neither recommends clinical use nor initiates synthesis; its outputs are computational hypotheses of open causal status. Human review is invoked at three checkpoints (non-redesignable or insufficient-evidence cases, conflicting precedents, and final candidate triage); every evidence item, precedent, edit, score, and human intervention is logged to the provenance memory. The aim is not to automate medicinal chemistry but to offer medicinal chemists an auditable testbed in which safety-driven redesign remains legible, contestable, and accountable, extensible to other precedent-guided design tasks.

Acknowledgements

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00431394) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Aronson, J. K. and Ferner, R. E. Joining the DoTS: new approach to classifying adverse drug reactions. *BMJ*, 327 (7425):1222–1225, 2003.

Averly, R., Baker, F. N., Watson, I. A., and Ning, X. Liddia: Language-based intelligent drug discovery agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 12015–12039, 2025.

Burley, S. K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G. V., Christie, C. H., Dalenberg, K., Di Costanzo, L., Duarte, J. M., et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research*, 49(D1):D437–D451, 2021.

Chandak, P., Huang, K., and Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. URL https://doi.org/10.1038/s41597-023-01960-3.

Chase, H. LangChain, October 2022. URL https://github.com/langchain-ai/langchain.

Edwards, I. R. and Aronson, J. K. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet*, 356 (9237):1255–1259, 2000.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

Hussain, J. and Rea, C. Computationally efficient algorithm to identify matched molecular pairs (mmps) in large data sets. *Journal of chemical information and modeling*, 50 (3):339–348, 2010.

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pp. 2323–2332. PMLR, 2018.

Kim, H., Moon, S., Zhung, W., Kim, S., Lim, J., and Kim, W. Y. Deepbioisostere: Discovering bioisosteres with deep learning for a fine control of multiple molecular properties. *arXiv preprint arXiv:2403.02706*, 2024.

Kramer, C., Fuchs, J. E., Whitebread, S., Gedeck, P., and Liedl, K. R. Matched molecular pair analysis: significance and the impact of experimental uncertainty. *Journal of Medicinal Chemistry*, 57(9):3786–3802, 2014.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

Liu, S., Lu, Y., Chen, S., Hu, X., Zhao, J., Lu, Y., and Zhao, Y. Drugagent: Automating ai-aided drug discovery programming through llm multi-agent collaboration. *arXiv preprint arXiv:2411.15692*, 2024.

Loeffler, H. H., He, J., Tibo, A., Janet, J. P., Voronov, A., Mervin, L. H., and Engkvist, O. Reinvent 4: Modern ai–driven generative molecule design. *Journal of Cheminformatics*, 16(1):20, 2024.

Park, B. K., Boobis, A., Clarke, S., Goldring, C. E., Jones, D., Kenna, J. G., Lambert, C., Laverty, H. G., Naisbitt, D. J., Nelson, S., et al. Managing the challenge of chemically reactive metabolites in drug development. *Nature Reviews Drug Discovery*, 10(4):292–306, 2011.

Rautio, J., Meanwell, N. A., Di, L., and Hageman, M. J. The expanding role of prodrugs in contemporary drug design and development. *Nature reviews drug discovery*, 17(8):559–587, 2018.

Ray, A. S., Fordyce, M. W., and Hitchcock, M. J. Tenofovir alafenamide: a novel prodrug of tenofovir for the treatment of human immunodeficiency virus. *Antiviral research*, 125:63–70, 2016.

Swanson, K., Walther, P., Leitz, J., Mukherjee, S., Wu, J. C., Shivnaraine, R. V., and Zou, J. Admet-ai: a machine learning admet platform for evaluation of large-scale chemical libraries. *Bioinformatics*, 40(7):btae416, 2024.

Tanaka, Y., Chen, H. Y., Belloni, P., Gisladottir, U., Kefeli, J., Patterson, J., Srinivasan, A., Zietz, M., Sirdeshmukh, G., Berkowitz, J., et al. Onsides database: Extracting adverse drug events from drug labels using natural language processing models. *Med*, 6(7), 2025.

Trott, O. and Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

UniProt Consortium, T. Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 46(5):2699–2699, 2018.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

You, J., Liu, B., Ying, Z., Pande, V., and Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

A. Appendix

A.1. Positioning Relative to Prior Agentic Systems

Recent agentic drug-discovery systems share one execution pattern: an LLM orchestrator retrieves data, invokes generators, and filters candidates with *in silico* tools. DrugAgent (Liu et al., 2024) automates the machine-learning programming behind discovery tasks. LIDDIA (Averly et al., 2025) coordinates the *in silico* stages of hit identification and lead optimization. Both frame the objective as a search over chemical space toward a target and a property profile. PRECEDE reuses this orchestration pattern, yet its contribution rests on three departures from prior work.

Problem formulation. Prior systems take a target and a property specification as input and return molecules that bind the target. PRECEDE takes an approved drug and one documented side effect as input and returns revisions that remove the liability while the therapeutic mechanism stays intact. The task is safety-driven redesign of an existing drug rather than *de novo* generation.

Precedent-guided reasoning. These systems optimize property scores for each case in isolation. PRECEDE abstracts a strategy from structured records of how medicinal chemists resolved analogous liabilities and transfers that strategy by analogy. The precedent memory is a resource that property-score optimizers do not use.

Attribution before generation. Many prior systems follow a generate-and-filter pattern, in which candidate molecules are proposed before downstream property and safety filters are applied. PRECEDE classifies the side effect by mechanism before any structural edit and admits only redesignable liabilities to redesign. Target-mediated effects halt at human review. The gate decides whether redesign is scientifically defensible before the generators run.

The contribution rests on the grounds of the reasoning rather than on the orchestration mechanics. The external tools are invoked under fixed execution settings, while the agentic component lies in non-trivial decisions: attribution classification, precedent selection under conflict, the reflect–refine–escalate loop, and rationale generation. Each decision is grounded in external evidence and recorded in the provenance log. The exception cases in Appendix A.5 (warfarin and terfenadine) demonstrate that the pipeline branches and rejects inputs rather than executes a fixed batch.

A.2. Pilot Evaluation Protocol

Table 1 summarizes the four evaluation axes introduced in Section 3. It specifies the metrics, data sources, and target criteria used to operationalize evidence grounding, precedent retrieval, redesign quality, and expert review in the initial PRECEDE pilot.

Support recall@k measures the fraction of held-out drug–side-effect associations for which relevant supporting evidence appears in the top-k results. **Citation validity** measures the fraction of cited sources that resolve to a verifiable record in SIDER, PrimeKG, PubMed, or the corresponding source database. **Attribution accuracy** is computed on a manually labeled subset of cases with literature-documented side-effect mechanisms; it tests whether PRECEDE assigns each case to the correct category among the five categories in Section 2. **Strategy-level F1** maps each retrieved precedent to an abstracted strategy class (Section 2) and compares it with the strategy documented in the held-out case. **Toxicity-proxy** Δ measures the relative change in predicted toxicity between parent and candidate compounds, averaged across multiple ADMET predictors to reduce dependence on a single model. **Expert review** uses three independent raters to score rationale plausibility on a 1–5 Likert scale. Inter-rater agreement is reported using Cohen's κ or an appropriate multi-rater agreement statistic.

Candidate success criterion. A candidate counts as a successful redesign only when it satisfies three conditions jointly. First, the therapeutic efficacy proxy stays within

Table 1. Evaluation axes, metrics, data sources, and target criteria for the PRECEDE pilot study.

Axis	Metric	Source	Target
Evidence grounding	support recall@k, citation validity, attribution accuracy	SIDER, PrimeKG, labeled subset	high agreement with curated associations
Precedent retrieval	top-k accuracy, strategy-level F1	held-out redesign cases	exceeds similarity-only baselines
Redesign quality	validity, SA, parent Tanimoto, docking Δ, toxicity-proxy Δ	RDKit, AutoDock Vina, ADMET predictors	reduced toxicity proxy with preserved efficacy proxy
Expert review	rationale plausibility (1–5), blinded, multi-rater	held-out redesign cases	mean ≥ 3.5, substantial agreement

Table 2. Seed precedent memory used by PRECEDE in the current proof-of-concept implementation.

Parent → successor	Targeted side effect	Strategy and structural change	Attribution
TDF → TAF	nephrotoxicity; bone mineral density loss	exposure modulation via phosphonamidate prodrug conversion	Exposure / PK
cisplatin → carboplatin	nephrotoxicity	liability-group replacement via cyclobutanedicarboxylate leaving group	Exposure / PK
carbamazepine → oxcarbazepine	hypersensitivity (rash)	keto-analog strategy via 10-keto substitution	Metabolism / reactive
terfenadine → fexofenadine	QT prolongation; torsades (hERG block)	active-metabolite switch via carboxylic-acid active metabolite	Off-target structural

Table 3. Knowledge resources and computational tools used across task modules in the implemented pipeline.

Task module	Knowledge resources and tools
Evidence grounding	SIDER, OnSIDES
Target resolution	ChEMBL, UniProt, RCSB PDB
SMILES generation	mmpdb, REINVENT Mol2Mol
Property prediction	ADMET-AI, RDKit
Binding / efficacy	AutoDock Vina

the parent-relative margin. Second, the targeted safety endpoint improves relative to the parent. Third, no non-targeted safety endpoint degrades beyond the parent-relative margin. The third condition screens the full ADMET-AI (Swanson et al., 2024) panel to flag cases where a redesign may trade the targeted liability for a new one.

A.3. Pilot Benchmark Curation

The historical replay benchmark is designed around 30–50 safety-driven redesign trajectories from two complementary sources. The first is curated medicinal chemistry case studies, including prodrug conversion and bioisosteric replacement (Rautio et al., 2018; Kim et al., 2024). An example is the transition from tenofovir disoproxil fumarate to tenofovir alafenamide. The second source is matched molecular pair literature derived from public bioactivity databases such as ChEMBL (Gaulton et al., 2012), where structural transformations associated with reduced off-target activity, such as hERG or CYP inhibition, have been systematically characterized (Kramer et al., 2014). Each trajectory contains a parent compound, a documented side effect, a structural modification, and a successor compound, with the modification rationale recorded as a ground-truth strategy label.

The pilot is kept deliberately modest in scale because the value of historical replay depends on the verifiability of each trajectory rather than the size of the corpus. Trajectories with ambiguous attribution, undocumented design rationale, or insufficient outcome data are excluded. We anticipate that scaling beyond the pilot will require automated mining of medicinal chemistry literature, which we treat as future work conditional on the validation of the pilot protocol.

A.4. Prototype Implementation Details

PRECEDE runs on a LangGraph orchestration graph (Chase, 2022) with conditional edges and a bounded reflection loop. The LLM is restricted to attribution, reflection, and rationale generation, while external tools perform retrieval, generation, scoring, and filtering under fixed execution settings. The backbone is Qwen3-30B-A3B-FP8 (Yang et al., 2025). Table 3 lists the tool components per task module. Table 2 lists the seed precedent memory used in the current proof-of-concept implementation.

A.5. Prototype Workflow Walkthrough

This section traces the implemented pipeline on a single case that the precedent memory does not contain, to demonstrate the control flow of PRECEDE rather than to validate the biological validity of the generated analog. The case separates strategy transfer from retrieval of a documented successor. The parent drug is cidofovir, a nucleotide analog with well-documented renal toxicity, and the target side effect is renal failure. The precedent memory holds no successor of cidofovir; the case therefore tests transfer of an abstracted strategy rather than retrieval of a stored answer.

Evidence grounding. The evidence module queries the side-effect resources SIDER and OnSIDES (Tanaka et al., 2025). Both resources support the cidofovir–renal failure association through two independent sources, with the matched terms “Renal failure” and “Renal failure acute”. The association meets the confidence threshold; the case proceeds to attribution.

Attribution-aware routing. The orchestrator assigns the exposure or pharmacokinetic (PK) category among the five categories of Section 2. The recorded rationale attributes the toxicity to renal tubular accumulation rather than to an on-target or off-target interaction. The PK category is redesignable; the case advances to precedent retrieval.

Precedent retrieval and strategy transfer. The retrieval module ranks precedents by attribution-category match and

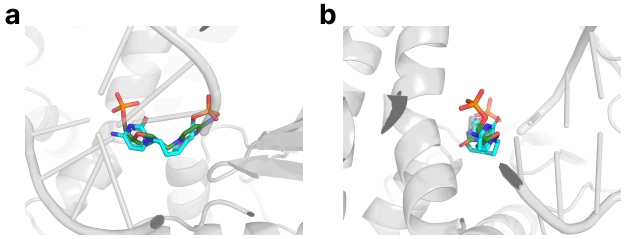

Figure 2. Docked poses on viral DNA polymerase (PDB 9JA3). Green: cidofovir (parent); cyan: generated analog.

by drug and effect token overlap. The two top precedents are the transition from tenofovir disoproxil fumarate to tenofovir alafenamide (TDF → TAF) and from cisplatin to carboplatin. Both precedents share the PK category with the query; neither precedent is a documented redesign of cidofovir. The module abstracts the strategy *exposure modulation* from TDF → TAF and transfers it to cidofovir.

Strategy-guided redesign. The redesign module generates candidates from the parent structure with two generators: matched molecular pair (MMP) (Hussain & Rea, 2010) transformations and a deep generative model (REINVENT Mol2Mol) (Loeffler et al., 2024). The module retains 23 valid candidates, 17 from REINVENT Mol2Mol and 6 from MMP transformations, after a chemical-validity filter and the removal of duplicates and parent-identical structures.

***In silico* evaluation.** The evaluation module resolves a therapeutic target of cidofovir to the herpes simplex virus type 1 (HSV-1) DNA polymerase (UniProt accession P04293) (UniProt Consortium, 2018) and selects Protein Data Bank (PDB) structure 9JA3 (Burley et al., 2021), a complex of the HSV-1 polymerase with duplex DNA and a bound nucleotide-analog inhibitor, for docking. The module predicts toxicity with ADMET-AI and selects microsomal intrinsic clearance (`Clearance_Microsome_AZ`) as the primary safety endpoint for the PK category. A higher clearance value indicates lower systemic exposure; higher values are treated as more favorable for this endpoint. The module ranks the 23 candidates by the docking-independent score components (safety endpoint delta, parent Tanimoto similarity, and synthetic accessibility). It then docks the parent and the three top-ranked candidates with AutoDock Vina (Trott & Olson, 2010) (Figure 2). The parent docking score is -7.88 kcal/mol. For these three docked candidates, docking-score deltas span -0.30 to $+0.24$ kcal/mol; all three remain within the parent-relative efficacy margin of 2.0 kcal/mol. The reflection module evaluates the top candidate and returns an accept decision (efficacy preserved, scaffold similarity maintained), terminating the loop.

Ranking and report. The system ranks candidates with a transparent weighted score over four normalized components: the primary safety endpoint delta (weight 0.45), docking efficacy (weight 0.30), parent Tanimoto similarity (weight 0.15), and synthetic accessibility (SA, weight 0.10). The score is normalized over the components available for each candidate. Table 4 summarizes the rank-1 candidate, which originates from REINVENT Mol2Mol, has a parent Tanimoto similarity of 0.83, and reaches a composite score of 0.94. The candidate docking score is -8.18 kcal/mol, a delta of -0.30 relative to the parent. The generated rationale frames the candidate as consistent with an exposure-modulation hypothesis derived from the transferred precedent strategy. Microsomal clearance is a hepatic measure and serves as an indirect proxy for renal exposure; this limitation motivates the human-supervised positioning of the system. The workflow terminates with a final human-review checkpoint for expert triage that marks every candidate as a hypothesis rather than a clinical or synthesis recommendation.

Table 4. Rank-1 candidate for the cidofovir walkthrough. The candidate originates from a strategy transferred from the TDF → TAF precedent.

Quantity	Value
Attribution category	exposure or pharmacokinetic
Transferred strategy	exposure modulation
Parent Tanimoto similarity	0.83
Primary safety endpoint	`Clearance_Microsome_AZ`
Safety endpoint delta	$+3.65$ (improvement)
Docking score (parent)	-7.88 kcal/mol
Docking score (candidate)	-8.18 kcal/mol
Efficacy preserved	yes
Composite score	0.94

Exception cases. Two cases test the first human-review checkpoint, where non-redesignable or insufficient-evidence inputs should halt before structural editing. The first case pairs warfarin with bleeding. The orchestrator assigns the target-mediated category because bleeding extends the intended anticoagulant mechanism rather than an off-target or pharmacokinetic liability. Because the target-mediated category is non-redesignable, the case halts at a human-review checkpoint without structural editing. The second case pairs terfenadine with QT prolongation. Although terfenadine appears in the seed precedent memory, the evidence module does not allow it to substitute for external side-effect support. Because the queried association is not covered by SIDER or OnSIDES in the current implementation, the case is assigned to the insufficient-evidence category and halts at a human-review checkpoint. Both cases illustrate that evidence grounding and attribution routing gate the pipeline as specified. Neither query reaches redesign.

A.6. Governance and Risk Considerations

Data provenance. PRECEDE relies on public research resources, including SIDER, PrimeKG, ChEMBL-derived

matched molecular pair datasets, and published medicinal chemistry case studies. The system does not ingest patient-level data or proprietary datasets in the proposed pilot.

Dual-use considerations. Side-effect mitigation strategies could, in principle, inform the inverse problem of increased side-effect risk. PRECEDE mitigates this risk by restricting outputs to literature-grounded modifications and abstracted strategies, and by logging generated candidates to the provenance memory. The system uses REINVENT only in parent-conditioned Mol2Mol mode and does not perform *de novo* toxicophore exploration or autonomous synthesis planning.

Human oversight protocol. The three review checkpoints described in Section 4 are operationalized as follows. First, cases routed to non-redesignable or insufficient-evidence categories are presented to a human reviewer with the supporting evidence and routing justification. Second, conflicting precedents above a disagreement threshold trigger human triage, where alternative hypotheses and contextual similarity scores are surfaced for selection. Third, candidate triage requires expert review of ranked candidates against the auditable redesign report before any downstream use.

Limitations. PRECEDE depends on public side-effect databases, which can under-report long-tail effects and encode reporting biases. Predicted toxicity proxies may exhibit domain shift on novel scaffolds. Historical replay agreement indicates consistency with documented medicinal chemistry logic, but it does not establish causal validity for novel redesign cases. These limitations motivate the human-supervised positioning of PRECEDE rather than autonomous deployment.