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

Anonymous Authors[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 adverse 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 structured precedents of prior safety-driven optimization, coordinated by an LLM orchestrator with explicit decision policies and human checkpoints. We position PRECEDE as a human-supervised AI-for-science workflow whose hypotheses remain auditable, falsifiable, and bounded by what prior pharmacology already supports.

1. Motivation and Task Definition

Many drug candidates exhibit therapeutic activity yet carry adverse effects that require further optimization 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 illustrated by prodrug optimization cases such as the transition from *tenofovir disoproxil fumarate* to *tenofovir alafenamide* (Ray et al., 2016), in which specific 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 (Liu et al., 2024; Averly et al., 2025). PRECEDE complements this line of work by grounding redesign in prior safety-driven evidence rather than property-score optimization.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026).

We define *side-effect-aware drug redesign* as follows. Given a parent drug x, a target adverse 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 causal screening of the adverse effect before any structural edit: the system must verify 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 verifies drug–side effect associations against SIDER (Kuhn et al., 2016) and biomedical knowledge graphs such as PrimeKG (Chandak et al., 2023), classifies the adverse 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 generated 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 adverse effect, and (iii) a historical replay protocol that tests whether the system's reasoning 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 adverse effect, and an optional therapeutic context, the orchestrator first verifies external support for the queried drug–side effect association.

Attribution-aware routing. Before redesign, PRECEDE classifies the adverse effect into one of five attribution categories based on established mechanism-based adverse-effect

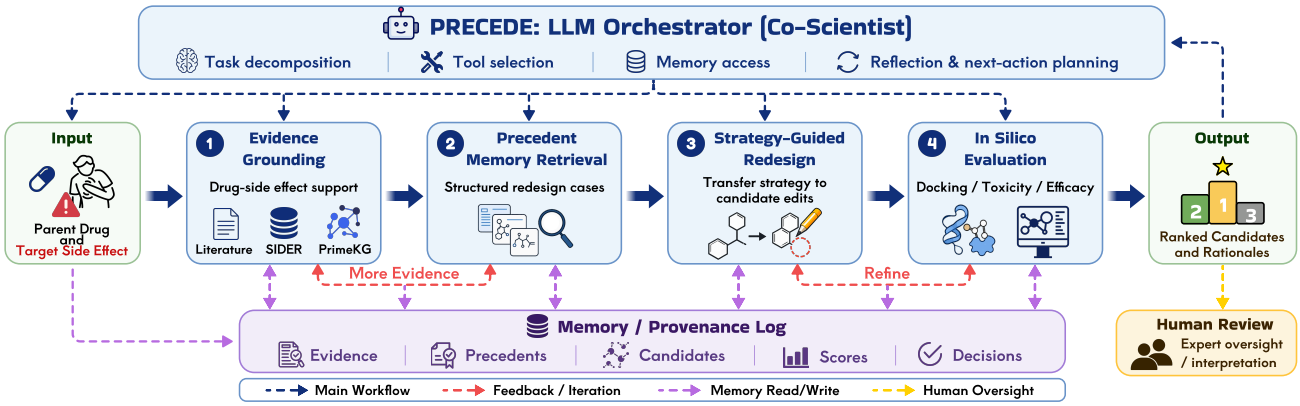

Figure 1. Overview of PRECEDE, an LLM-orchestrated workflow for evidence-grounded, precedent-guided drug redesign.

taxonomies (Edwards & Aronson, 2000; Aronson & Ferner, 2003; Park et al., 2011): (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 structural redesign; (i) triggers alternative-target hypothesis generation, and (v) is flagged for human review. This routing makes structural editing scientifically defensible and constrains the search 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 whose precedents match 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. Explicit policies govern transitions between stages: evidence below a confidence threshold (*e.g.*, absence of SIDER-confirmed association or fewer than two supporting citations) triggers retrieval expansion; conflicting precedents are resolved by ranked alternative hypotheses scored on contextual similarity; failed evaluation (*e.g.*, docking degradation beyond a parent-relative margin, or low synthetic accessibility) returns the system to redesign with tightened constraints. A provenance log records every evidence item, precedent, edit, score, and decision rationale, and exports an auditable redesign report at termination.

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 and 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 agreement 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 compound with its safety-driven successor. Holding out the successor, PRECEDE must recover the documented optimization from the parent and target adverse effect alone. We treat replay as necessary but not sufficient: agreement supports reasoning consistent with medicinal chemistry, while interpretable disagreement remains informative.

4. Governance and Expected Impact

PRECEDE is positioned as a human-supervised hypothesis-generation system rather than an autonomous clinical or synthesis engine: it neither recommends clinical use nor initiates synthesis, and its outputs are computational hypotheses whose causal status remains open. Human review is invoked at three checkpoints (insufficient-evidence cases, conflicting precedents, and final candidate triage), and every evidence item, precedent, edit, score, and human intervention is logged to the provenance memory for attribution and audit. The aim is not to automate medicinal chemistry but to provide a testbed in which the reasoning behind safety-driven redesign remains legible, contestable, and accountable to humans who remain responsible for the science.

References

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

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.

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

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.

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.

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. Evaluation Details

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 adverse-effect mechanisms; it tests whether PRECEDE assigns the correct category among the five attribution classes 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.

A.2. Pilot Benchmark Curation

The historical replay benchmark consists of 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). A representative 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, 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 adverse effect, structural modification, and 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 corpus size. 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 validation of the pilot protocol.

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

A.3. 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 increasing adverse-effect risk. PRECEDE mitigates this risk by restricting outputs to literature-grounded modifications and abstracted strategies, and by logging all generated candidates to the provenance memory. The system 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 adverse-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.