SureRoute: Toward a Hallucination-Free Self-Improving Platform for Retrosynthesis

Jieli Zhou **Naiwu Chen** **Longzhang Liu** **Peiyu Zhang***
AI Chemistry Group, Future Chemistry Department, XtalPi Inc
{jieli.zhou, naiwu.chen, longzhang.liu, peiyu.zhang}@xtalpi.com

Abstract

AI models, including large language models, are increasingly integrated into scientific discovery workflows, yet they remain prone to hallucination. In experimental sciences, such errors translate directly into failed wet-lab validations and wasted resources; in self-improving agentic systems, confident errors risk being reinforced rather than corrected. Retrosynthesis provides a representative example of this failure mode: existing models can generate chemically plausible routes, but cannot reliably determine which routes are experimentally feasible. We define **Chemical Hallucination** as a route that appears valid yet fails under competing reactive sites, unresolved selectivity, or missing mechanistic support, a failure largely invisible to the Recall@K metric. We introduce **SureRoute**, a chemical verifier-anchored retrosynthesis platform that suppresses Chemical Hallucination. SureRoute combines a multi-model ensemble, data asset retrieval, and **ChemHarness**, an executable chemical intuition engine for route verification and reliability-first ranking. On a benchmark of 350 real-world industrial targets, SureRoute reaches 74.3% recall@1, 2.2–3.5$\times$ that of seven single-step models and three frontier LLMs, while cutting top-1 Chemical Hallucination to 4.6%, a 4–6$\times$ reduction relative to frontier LLMs. As a model-agnostic reranker, ChemHarness drives detectable hallucination toward near-zero across arbitrary backbone candidates. SureRoute shows that reliable scientific AI requires not only strong generation, but executable verification.

1 Introduction

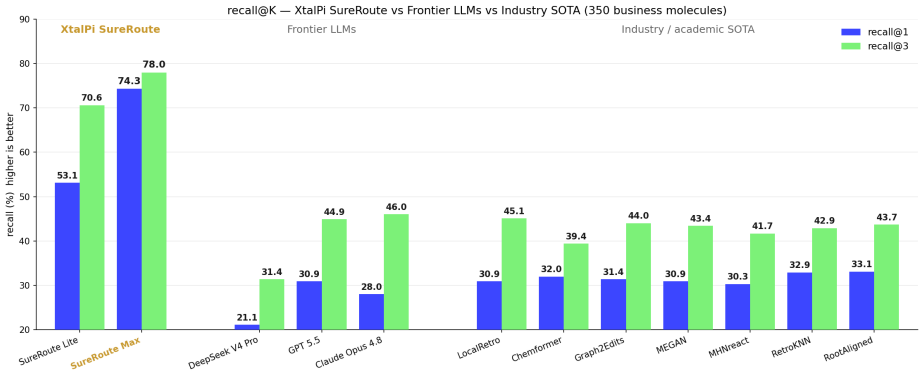


Figure 1: recall@1/@3 across 7 single-step models, 3 frontier LLMs, and SureRoute (Lite: ensemble; Max: full system).SureRoute's recall@1 is a step-function jump above every other method.

Hallucination—fluent but factually unsupported generation—is a central obstacle to deploying large language models in high-stakes domains [Ji et al., 2023, Huang et al., 2025, Zhang et al., 2025c]. In dialogue, a hallucination fact costs a correction; in scientific discovery, it costs an experiment: an infeasible synthesis route consumes real laboratory resources before the error is exposed. Models are also poor judges of their own errors—intrinsic self-correction without external feedback often fails to improve, and can even degrade reliability [Huang et al., 2024]. The risk is amplified in *self-improving* systems, where models generate their own training signal through bootstrapped reasoning, self-refinement, or self-rewarding loops [Zelikman et al., 2022, Madaan et al., 2023, Yuan et al., 2024]: hallucinations become training data, and confident errors are reinforced across iterations. What such loops lack in scientific domains is an *executable verifier*—an external harness that checks proposals against domain rules rather than the model's own beliefs.

We study a scientific-discovery setting where this problem is both acute and measurable: **retrosynthesis**. Single-step retrosynthesis is the atomic decision from which every multi-step route is built [Segler et al., 2018, Chen et al., 2020]; if it is wrong, the search tree expands around a false premise and a plausible-looking route may be mistaken for an executable plan. Yet for a decade the field has optimized for reference matching rather than chemical validity: the dominant protocol asks whether top-K predictions match reference reactants in curated benchmarks such as the USPTO series [Lowe, 2012, Liu et al., 2017, Tetko et al., 2020]. Models optimized for this metric, including recent LLM-based systems, propose disconnections that are syntactically and superficially plausible—but the metric never asks whether the transformation would work on *this substrate*: whether selectivity is controllable, whether competing reactive sites exist, whether the disconnection has mechanistic or literature precedent.

We call routes that fail on this second axis **Chemical Hallucinations**: superficially valid disconnections that an experienced chemist would reject for unresolved competing reactive sites or lack of mechanistic basis. The analogy to LLM hallucination is direct—fluent, well-formed, and false—except the violated constraint is organic chemistry: selectivity, mechanism, and precedent. This failure is largely invisible to recall@K: a model can match a literature reactant set while ranking chemically dangerous alternatives highly, and can improve recall without avoiding selectivity traps. In our evaluation, scaling does not remove this failure mode; frontier LLMs, despite strong chemical fluency, often produce confident, professionally phrased, and unexecutable routes.

The stakes extend beyond single predictions. A growing body of work retrains the single-step policy on the system's own search results, distilling successful routes back into the model in the spirit of expert iteration and AlphaGo Zero-style self-play [Anthony et al., 2017, Silver et al., 2017, Kim et al., 2021, Liu et al., 2023, Guo et al., 2024]. This reduces dependence on human-labeled data, but the loop is only as trustworthy as the *verifier* deciding which self-generated routes to learn from. Wet-lab validation is trustworthy but slow and expensive; letting a model score its own output reintroduces exactly the failure this paper targets. A route that merely *looks* sound can be graded a success and amplified from an isolated mistake into a systematic prior—a concrete instance of *misevolution* in self-improving agents [Shao et al., 2026], made more dangerous here because the error is chemical rather than syntactic.

This motivates our central question: **can a verifier approach wet-lab trustworthiness while retaining the speed and scalability of computational evaluation?** ChemHarness is our answer: an executable chemical intuition engine that checks candidate routes for functional-group competition, regio-/chemoselectivity risk, mechanistic plausibility, and literature support. SureRoute organizes the entire platform around it as a reliability layer: generation proposes candidates, verification filters hallucinations, ranking prioritizes verified routes, and an evolution layer converts expert-confirmed failures into new ChemHarness data (details in a companion technical report).

We position SureRoute not as another single-step predictor, but as a verifier-anchored self-improving platform for reliable retrosynthesis. This report makes three contributions:

1. **SureRoute**, a four-layer platform combining a multi-model intelligent ensemble, literature-precedent retrieval, the ChemHarness chemical intuition engine, and an evolution layer that learns from expert-confirmed failures.

2. **A reliability-focused evaluation** against seven state-of-the-art single-step models and three frontier LLMs on 350 real-world industrial molecules, measuring both recall@K and Chemical Hallucination rate under a unified protocol.

3. **A demonstration that ChemHarness is model-agnostic and pluggable**: applied post-hoc to any backbone's top-10 candidates, it drives detectable top-1 Chemical Hallucination toward near-zero, including for LLMs never tuned for it.

2 Related Work

We organize related work into four threads: single-step retrosynthesis prediction (Section 2.1), multi-step planning and search (Section 2.2), self-improving retrosynthesis (Section 2.3), and the broader self-evolution and verification-bottleneck literature in scientific AI (Section 2.4). The first two threads sketch the specialized technology stack SureRoute builds on; the latter two frame where SureRoute's core contribution sits.

2.1 Single-step retrosynthesis models

Single-step models are commonly grouped by whether they rely on explicit reaction templates. **Template-based** methods score candidate transformations against a library of reusable rules [Segler and Waller, 2017, Dai et al., 2019]; LocalRetro predicts atom/bond-level local templates on the chemical intuition that reaction changes are mostly local, trading some generalization for interpretability and built-in chemical validity [Chen and Jung, 2021]. **Template-free** methods largely treat retrosynthesis as SMILES-to-SMILES sequence translation—Molecular Transformer-style seq2seq models with SMILES augmentation generalize well but can emit syntactically or chemically invalid output with no built-in feasibility constraint [Karpov et al., 2019, Tetko et al., 2020]. **Semi-template and graph-based** methods try to combine both strengths: RetroXpert first localizes the reaction center before completing the synthon, and graph-edit formulations represent the transformation as a sequence of graph edits; more recent generative approaches such as Markov-bridge and discrete-flow-matching formulations push single-step diversity and accuracy further [Yan et al., 2020, Igashov et al., 2024, Yadav et al., 2026].

For SureRoute, single-step models are a pluggable **generation layer**: the platform is not tied to any one architecture, but pools multiple single-step models (template-based and template-free alike) behind a unified candidate-proposal interface, leaving organization and filtering to the search and verification layers described in Section 3.

Beyond dedicated architectures, general-purpose LLMs have recently been applied to retrosynthesis purely via prompting, exhibiting surprisingly strong disconnection recall without task-specific training. LLM retrosynthesis outputs, however, are typically evaluated the same way as specialized models—recall against a reference set—leaving open whether the model's *ranking* of candidates reflects real chemical risk. We treat frontier LLMs as an additional class of backbone under our unified evaluation protocol in Section 5, rather than assuming their fluency implies reliability.

2.2 Multi-step retrosynthesis planning and search

Above single-step prediction, a multi-step search algorithm determines overall route quality. Monte Carlo Tree Search brought neural-symbolic retrosynthetic planning toward expert-level search, while Retro* proposed AND-OR-tree best-first search guided by a learned value function; later graph-search and experience-guided variants further improved expansion reuse and pruning [Segler et al., 2018, Chen et al., 2020, Xie et al., 2022, Hong et al., 2023]. A notable recent thread is **planning under uncertainty**: Retro-fallback argues that real-world reaction feasibility is itself stochastic and only partially known, and optimizes for whether a route survives in the lab rather than purely on internal model metrics [Tripp et al., 2024]. Re-evaluation frameworks such as Syntheseus further emphasize the importance of unified, fair multi-step benchmarking, which directly informs our own benchmark design in Section 4 [Maziarz et al., 2025].

2.3 Self-improving retrosynthesis

The line of work most directly related to SureRoute feeds search results back into training, closing a self-improvement loop. Its lineage traces to Expert Iteration and AlphaGo Zero: a strong policy is distilled from the outcomes of its own search, then the strengthened policy searches better, and the

cycle repeats [Anthony et al., 2017, Silver et al., 2017]. Self-Improved Retrosynthetic Planning is an early retrosynthesis instance of this idea [Kim et al., 2021].

PDVN frames retrosynthesis as a tree-structured MDP and trains a policy with dual value networks (synthesizability and cost) directly on whether complete routes solve, rather than only on single-step accuracy [Liu et al., 2023]. ReSynZ more directly replicates the AlphaGo Zero recipe for template-based retrosynthesis—MCTS plus reinforcement learning self-improvement—and shows that a strong verifier can substitute for a large training corpus [Guo et al., 2024].

The introduction of large language models has pushed this line further. RetroDFM-R drives reasoning-style retrosynthesis with chemically verifiable rewards; Retro-R1 trains an LLM agent to call single-step tools over multiple turns via reinforcement learning; RTRL exploits round-trip consistency between forward and retro tasks to self-reinforce a chemical LLM without labeled data; and end-to-end chain-of-thought approaches internalize multi-step logic entirely inside a single reasoning model, removing the external search heuristic altogether [Liu et al., 2026, Zhang et al., 2025b, Zuo et al., 2026].

A clear trend runs through this body of work: reward signals are shifting from human annotation, through model self-evaluation, and converging on **verifiable rewards**. Several of the most recent LLM-based methods explicitly build "verifiability" into their core objective—which is exactly the point made in Section 1: verifier quality is the lever on which the whole loop turns. But the verifiability on offer in this literature is still almost entirely a **single-step** notion (does the predicted reactant set exactly match a reference, do atoms balance)—a **route-level, executable, competition-aware** notion of chemical trustworthiness remains largely unaddressed. This is precisely the gap SureRoute and ChemHarness target: verification is elevated from plausibility to executable verification, so that a self-improvement loop consumes a signal close to wet-lab semantics rather than a model's own optimistic self-assessment.

2.4 Self-evolution and the verification bottleneck in scientific AI

Zooming out from retrosynthesis to scientific AI broadly, recent surveys organize self-evolving agents into four carrier paths [Gao et al., 2026]: **weight evolution** (a model updates its own parameters from self-generated data), **memory evolution** (parameters stay frozen while an external experience store grows), **tool/code evolution** (the evolving artifact is executable code), and **physical closed-loop evolution** (wet-lab experiments serve as the ultimate verifier). Self-improving retrosynthesis (Section 2.3) sits mainly on the weight-evolution path.

On the **memory-evolution** path, ChemAgent lets an LLM improve with experience via self-updating planning/execution/knowledge memories without any parameter update [Tang et al., 2025]. On the **tool/code-evolution** path, FunSearch and AlphaEvolve show that evolving *programs* rather than answers lets the verifier be cheap and exact (a unit test), while the Darwin Gödel Machine and Gödel Agent let an agent rewrite its own code—ChemHarness's practice of compiling chemical knowledge into executable, rule-based checks is the chemistry-domain analogue of this path [Romera-Paredes et al., 2024, Novikov et al., 2025, Zhang et al., 2025a, Yin et al., 2025]. On the **physical closed-loop** path, autonomous experimentation systems demonstrate the most trustworthy verifier available—real reaction outcomes—but also the highest iteration cost; a recent survey of self-driving laboratories emphasizes that most deployed systems remain far from full autonomy [Boiko et al., 2023, Szymanski et al., 2023, Burger et al., 2020, Lee et al., 2026].

A shared bottleneck runs through all four paths, and it has recently shifted from "generation capability" to "verification capability." Work on emergent misevolution risk in self-improving agents makes this shift explicit from a safety angle [Shao et al., 2026]: self-evolution without a trustworthy verifier will systematically amplify bias and hallucination rather than correct it. **SureRoute's position responds to this from the engineering side rather than pushing verification all the way to the most expensive physical extreme**: it occupies a middle ground between weight-evolution loops and physical closed-loop verification—using auditable, executable computational verification (ChemHarness's competition detection and hallucination typing) as a cheap-but-trustworthy reward signal, and using a layered annotation pipeline to accumulate each cycle's failure cases as a compounding asset. Together, these give a retrosynthesis system that can both keep improving itself and resist misevolution while doing so.

3 Method

SureRoute is not a single model but a four-layered self-improving system. Each layer addresses a failure mode that the previous layer alone cannot solve (Figure 2, Table 1).

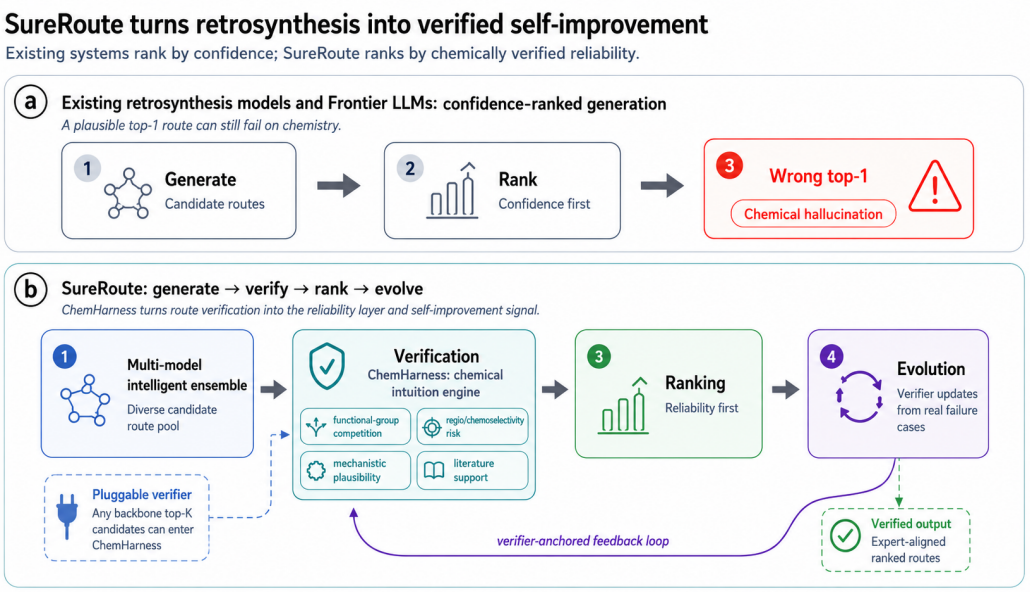


Figure 2: SureRoute converts retrosynthesis from confidence-ranked generation into verifier-anchored self-improvement. Existing retrosynthesis systems typically generate candidate routes and rank them by model confidence, which can surface a plausible but chemically hallucination top-1 route. SureRoute instead uses a four-layer architecture: a generation layer builds a diverse candidate pool; ChemHarness verifies each candidate for competition, selectivity, mechanism, and precedent; the ranking layer prioritizes chemically verified routes over confident hallucinations; and the evolution layer converts expert-confirmed failures into high-quality data. The verifier is pluggable: any backbone model's top-K candidates can enter the verification layer.

None of the four layers is sufficient alone: generation-only ensembles can still walk into selectivity hallucinations; retrieval alone is limited to known precedent; verification without a broad candidate pool has nothing to promote; and self-improvement without a trustworthy verifier risks reinforcing confident mistakes. SureRoute combines all four so the delivered top-1 route is selected by chemically verified reliability rather than raw model confidence.

Table 1: The four layers of SureRoute and the failure mode each addresses.

Layer	Component	Addresses
Generation	Multi-model retrosynthesis ensemble	Expands the candidate pool and reduces single-model blind spots across reaction families and scaffolds.
Verification	ChemHarness	Screens every candidate for functional-group competition, selectivity risk, mechanistic plausibility, and precedent support.
Ranking	Chemical-reliability-aware reranking	Promotes clean, precedent-supported routes above high-confidence hallucinations, changing top-1 selection from model confidence to verified feasibility.
Evolution	Expert-confirmed failure feedback loop	Converts rejected production routes into high-quality training data that strengthen the verifier over time.

3.1 Multi-model ensemble candidate pool

For each target molecule, we independently run multiple SOTA single-step retrosynthesis models and collect each model's top-10 ranked candidate disconnections. We develop a novel model ensemble strategy to combine these disconnections. On a high-level, ensemble follows the reciprocal-rank-fusion principle used in information retrieval [Cormack et al., 2009]. Because different architectures (template-based, graph-edit, retrieval-augmented, and transformer-based) fail on different substrates, we pool candidates from all models and aggregate rank information via **Reciprocal Rank Fusion (RRF)**: each unique (canonicalized) reactant set receives a score

$$s(c) = \sum_{m \in \text{models}} \frac{1}{k_c + \text{rank}_m(c)}$$

where $\text{rank}_m(c)$ is the rank the candidate received from model m (candidates not proposed by a model do not contribute), and k_c is a constant damping the influence of low-rank positions. This produces a single deduplicated, RRF-ranked candidate pool per target that is broader than any single model's coverage—a purely rank-fused ensemble (SureRoute Lite) already greatly improves recall@1 over the strongest single model (Section 5.1), before any reliability layer is applied.

Candidates are compared and deduplicated after a canonicalization step (Section 4) that strips reagents/solvents/salts from both the predicted and reference sides and reduces each candidate to a set of RDKit InChIKeys, so that superficially different SMILES representing the same reactant set are correctly merged [RDKit contributors, 2024].

3.2 ChemHarness: chemical reliability verification

ChemHarness is treated in this report as the verification layer of SureRoute. The details of ChemHarness will be described in a separate technical report. Here, we focus on the system-level question: given a candidate pool from modern retrosynthesis models and LLMs, can an executable chemical verifier reliably demote hallucination routes and improve top-ranked retrosynthetic suggestions?

At a high level, ChemHarness evaluates each candidate disconnection along different reliability dimensions.

Competition and selectivity risk. ChemHarness checks whether the proposed route contains unresolved functional-group competition or regio-/chemoselectivity ambiguity. These risks arise when a substrate contains multiple chemically plausible reactive sites, and the model proposes a disconnection without explaining why the desired site should dominate. Rather than trusting the model's confidence score, SureRoute uses ChemHarness to identify such selectivity hallucinations and prevent them from being over-ranked.

Mechanistic plausibility. ChemHarness also checks whether the proposed transformation is consistent with the expected chemistry of the substrate and reaction class. This targets a common failure mode of both single-step models and LLMs: proposing a formally balanced disconnection that lacks a viable mechanistic basis. Candidates that fail this check are treated as chemically unreliable even if they appear syntactically valid.

Precedent support. In addition to internal chemical checks, ChemHarness estimates whether a candidate route is supported by relevant literature or database precedent using an intelligent similarity metric. This signal is not used as a blind lookup answer, but as supporting evidence for route reliability: a route with close precedent is favored over a route that is both mechanistically uncertain and unsupported by known chemistry.

The output of ChemHarness is a compact reliability profile for each candidate route, indicating whether the route is likely to be competition-free, mechanistically plausible, and precedent-supported. SureRoute then uses this profile for reliability-first reranking: chemically verified candidates are promoted, while routes that look plausible but contain unresolved chemical risks are demoted.

This design deliberately separates *generation* from *verification*. The generation layer is encouraged to be broad and diverse, while ChemHarness acts as a model-agnostic reliability filter that can be applied to candidates from any backbone model. Because ChemHarness is executable and rule-based at the

system level, it can be updated through expert-confirmed failures without retraining the underlying generative models.

3.3 Reranking: layered candidate ordering

The final top-K route list served to the user is produced by a layered sort, most-important criterion first:

1. **Literature/database hit**—if the ensemble's disconnection matches a real literature reaction for this exact target, that route is placed first with maximal confidence, since it is a route a chemist has demonstrably already executed.
2. **No functional-group competition**—among remaining candidates, routes ChemHarness has *not* flagged for selectivity risk are ranked above flagged ones, regardless of the underlying model's raw confidence score.
3. **Precedent score**—among otherwise-equal candidates, higher literature-precedent score is preferred; this is also used to diversify the second and third ranked slots so that routes 2–3 are not simply near-duplicates of route 1.
4. **Ensemble confidence (RRF score)**—the residual tie-breaker.

This ordering choice—competition-free before raw-confidence—is deliberate and is the central mechanism by which SureRoute converts ensemble recall into low-hallucination top-1 recall: a route that is individually the single most-confident prediction from the strongest model, but that ChemHarness flags as sitting on a competing site, is demoted below a less-confident but competition-free alternative.

3.4 Pluggability

ChemHarness's competition/plausibility checks and precedent scoring operate purely on a proposed reaction (reactants + product), independent of which model proposed it. This means the reranking procedure in Section 3.3 can be applied to the top candidates from *any* backbone—a different single-step model, or a frontier LLM prompted for retrosynthesis—with no retraining. Section 5.3 demonstrates this directly: applying ChemHarness reranking to LLM and single-step-model candidate pools drives their top-1 Chemical Hallucination rate toward 0%, even though ChemHarness's rules were developed independently of any of these backbones.

3.5 ChemHarness as the anchor for a self-improving loop

As motivated in Section 1, a self-improvement loop for retrosynthesis is only as trustworthy as its verifier. Wet-lab feedback is the gold standard but is far too slow and expensive to gate every candidate route; letting a model score its own output is fast but reintroduces the exact failure this paper documents, since a confidently wrong route can be graded as a success and distilled back into the system. ChemHarness is designed to sit between these two extremes: its competition and plausibility checks are executable and deterministic (Section 3.2), so a verdict can be produced and audited in milliseconds rather than weeks, while still being anchored in explicit chemical rules rather than a model's self-assessment.

This property lets ChemHarness close a rule-improvement loop entirely within the computational domain. Every route SureRoute serves in production is a candidate observation: when a chemist overrides or rejects a top-ranked route, or when downstream forward-synthesis checks disagree with a route ChemHarness had marked clean, that disagreement is logged as a confirmed failure case rather than discarded. These confirmed failures are triaged into the same rule families described in Section 3.2 and used to either tighten an existing competition rule (adding a reactive-site pattern the rule previously missed) or carve out a new exception (a substrate pattern the rule previously over-flagged). Because the loop's reward signal is a rule violation checked against a real, chemist-confirmed outcome—not a model's own confidence—a route that merely *looks* mechanistically sound cannot be mistaken for a validated success and cannot be distilled back in as a false positive, which is the specific misevolution risk described in Section 2.4.

Two properties make this loop compound rather than plateau. First, coverage grows with usage: every production molecule SureRoute processes is a potential source of new failure cases, so the

rule set's blind spots shrink as deployment scales, without requiring a proportional increase in wet-lab or manual-annotation cost. Second, the loop is auditable at every step—because rules given by ChemHarness are explicit and human-readable rather than distributed across model weights, a chemist can inspect exactly which rule fired, agree or disagree with it, and correct it directly, which is not possible for a black-box learned reranker. This combination—cheap-but-trustworthy verification, plus a rule representation a domain expert can directly edit—is what makes SureRoute's reliability advantage as compounding: it is not a fixed asset delivered once, but a system property that keeps improving with production usage in a way a static, pretrained model cannot replicate without repeating the same accumulation process.

4 Benchmark: industrial retrosynthesis evaluation set

4.1 Motivation

Standard retrosynthesis benchmarks mainly evaluate whether a model's top-K predicted reactant sets match a reference reaction in curated public datasets such as USPTO [Lowe, 2012, Liu et al., 2017, Tetko et al., 2020, Maziarz et al., 2025]. This is useful for measuring disconnection recall, but it is not sufficient for measuring route reliability in industrial settings. In real projects, a route can look reasonable at the reactant-set level while still being unsuitable for execution. To evaluate this gap, we constructed an internal industrial retrosynthesis evaluation set.

4.2 Dataset

The evaluation set contains 350 real industrial molecules selected from practical retrosynthesis use cases. These molecules are more challenging than textbook examples: they contain realistic scaffolds, multiple functional groups, and substrate-specific constraints that require chemical judgment rather than pattern matching alone.

For each target, we collected expert-validated reference routes and compared model-generated candidates against these references under a unified evaluation protocol.

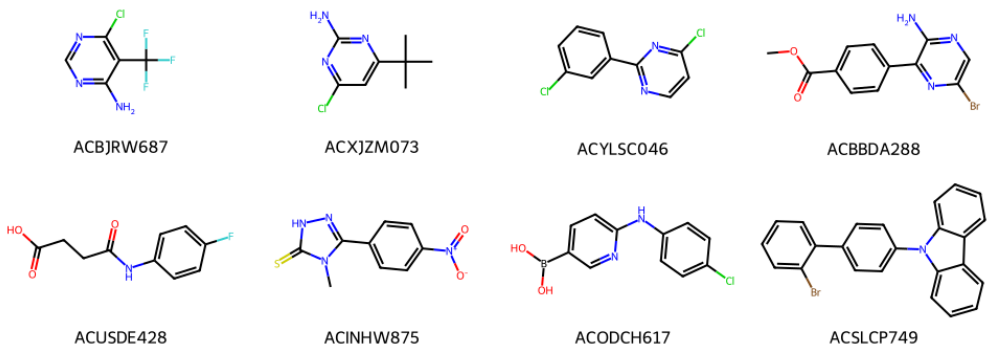


Figure 3: Representative molecules from the internal industrial evaluation set. The benchmark focuses on realistic industrial targets rather than simplified textbook reactions. Molecule IDs can be queried from the XtalPi Aifchem website https://www.aifchem.com/.

4.3 Evaluation protocol

Task. Given a target molecule, each system proposes up to 5 ranked single-step retrosynthetic candidate reactant sets.

Matching. A prediction is counted as a hit only when the predicted reactive-core reactant set matches the expert-validated reference after the same standardization procedure is applied to all methods. Reagents, solvents, catalysts, salts, and other non-core components are handled consistently across models so that the comparison measures retrosynthetic disconnection quality rather than formatting differences.

Chemical reliability. In addition to recall@K, we evaluate whether the top-ranked route contains chemical reliability risks. This allows us to measure not only whether a model can generate a reference-like answer, but also whether its top recommendation is chemically trustworthy. All routes are carefully inspected and annotated by an expert chemist.

Fair comparison. All single-step models, LLMs, and SureRoute variants are evaluated with the same targets, same candidate limit, same standardization pipeline, and same scoring criteria. Literature-retrieval matches are excluded from the main recall comparison when needed to avoid inflating performance through direct database overlap.

5 Experiments

5.1 Disconnection recall@K

Table 2: recall@K on the 350-molecule benchmark.

Method	recall@1	recall@3	recall@5
LocalRetro	30.9%	45.1%	49.1%
Chemformer	32.0%	39.4%	40.3%
Graph2Edits	31.4%	44.0%	49.4%
MEGAN	30.9%	43.4%	45.7%
MHNreact	30.3%	41.7%	44.3%
RetroKNN	32.9%	42.9%	46.0%
RootAligned	33.1%	43.7%	46.3%
DeepSeek V4 Pro	21.1%	31.4%	38.3%
GPT-5.5	30.9%	44.9%	50.9%
Claude Opus 4.8	28.0%	46.0%	54.0%
SureRoute Lite (model ensemble, no ChemHarness)	53.1%	70.6%	72.3%
SureRoute Max (model ensemble, with ChemHarness)	**74.3%**	**78.0%**	**78.6%**

Table 2 and Figure 1 report recall@K under the protocol in Section 4.3, for 7 SOTA single-step models, 2 SureRoute configurations, and 3 frontier LLMs. The specialized backbones represent the main template-based, template-free, graph-based, and retrieval/alignment-style families developed in the retrosynthesis literature [Segler and Waller, 2017, Chen and Jung, 2021, Karpov et al., 2019, Tetko et al., 2020, Yan et al., 2020].

SureRoute Max's recall@1 of 74.3% is 1.4× the plain-ensemble baseline (SureRoute Lite, 53.1%) and 2.2–3.5× every single-step model (30–33%) and every frontier LLM (21–31%). Because a chemist typically acts on the first proposed route, recall@1 is the most operationally relevant number in this table, and the gap here is far larger than at recall@3 or recall@5, where ensembling alone (without reliability reranking) already closes much of the distance. This is consistent with our thesis: the largest single lever for realistic top-1 recall is not more/better disconnection candidates but *correctly discarding the confident-but-wrong ones*—which is exactly what ChemHarness's reranking (Section 3.3) is designed to do, and which pure ensembling cannot achieve by itself.

5.2 Chemical Hallucination rate

Recall@K alone cannot distinguish "the model's top choice is chemically sound" from "the model's top choice happens to match the literature answer by chance." We therefore report **Chemical Hallucination rate**: the fraction of targets for which a method's *actual delivered top-1 route* has either functional-group competition or mechanistic implausibility. Lower is better; a route is counted as a hallucination if it fails *either* check. The inspections are carefully done by experienced XtalPi Chemists.

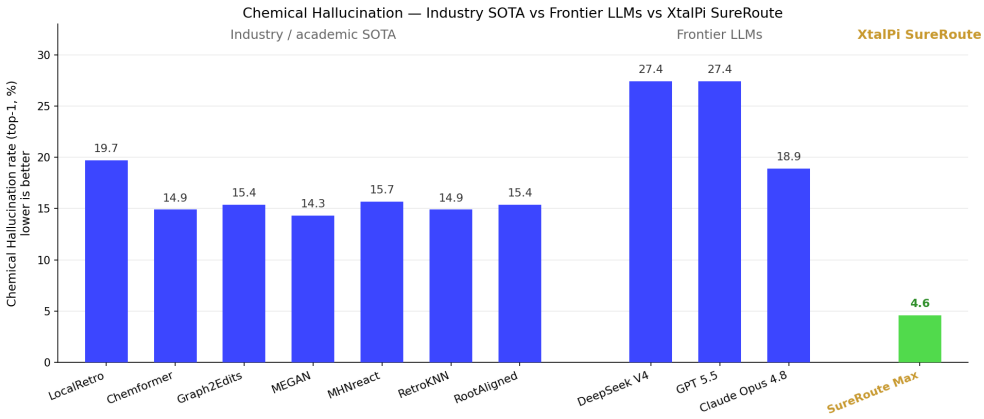


Figure 4: Top-1 Chemical Hallucination rate across all methods. Frontier LLMs (18.9–27.4%) are *worse* than every single-step model (14.3–19.7%) on this axis despite matching or exceeding them on raw recall (Figure 1)—fluent disconnection does not imply selectivity awareness. SureRoute's 4.6% is the only bar in the single digits.

All 7 single-step models hallucinate at 14–20% on their top-1 route (Figure 4, Table 3). The three frontier LLMs are *worse*, not better, at 18.9–27.4%: fluency and disconnection accuracy do not transfer into selectivity awareness. SureRoute's top-1 Chemical Hallucination rate of 4.6% is achieved by combining literature-route injection with ChemHarness-based reranking (Section 3.3), which actively demotes competition-flagged candidates rather than merely reporting whichever candidate scored highest under the underlying model's own confidence.

Table 3: Chemical Hallucination rate (top-1 route).

Method	Chemical Hallucination (top-1)
LocalRetro	19.7%
Chemformer	14.9%
Graph2Edits	15.4%
MEGAN	14.3%
MHNreact	15.7%
RetroKNN	14.9%
RootAligned	15.4%
DeepSeek V4	27.4%
GPT-5.5	27.4%
Claude Opus 4.8	18.9%
SureRoute Max	**4.6%**

5.3 Pluggability: reranking arbitrary backbones with ChemHarness

To test whether ChemHarness's benefit is specific to SureRoute's own ensemble or generalizes to arbitrary backbones, we take each method's top-10 raw candidates (single-step models and LLMs alike) and rerank them by ChemHarness's competition/plausibility signal alone (Section 3.3, criterion 2), with no other change to the candidate pool (Figure 5, Table 4).

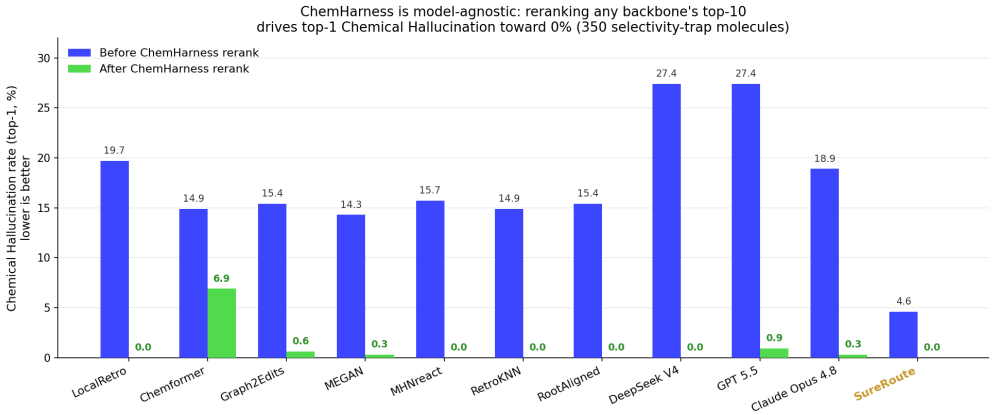


Figure 5: Top-1 Chemical Hallucination rate before (blue) and after (green) applying ChemHarness reranking to each backbone's own top-10 candidates, with no other change to the underlying model. Every bar collapses toward zero, including for LLMs ChemHarness was never tuned against.

Reranking collapses ChemHarness-detectable top-1 Chemical Hallucination toward near-zero for every backbone we tested, specialized model and frontier LLM alike, despite ChemHarness's rules having been developed without reference to any of these specific models. This is the central evidence for our pluggability claim: the reliability layer is separable from the disconnection-generation layer, and can be composed with a system it was never tuned against.

Table 4: Top-1 Chemical Hallucination rate, before → after ChemHarness reranking.

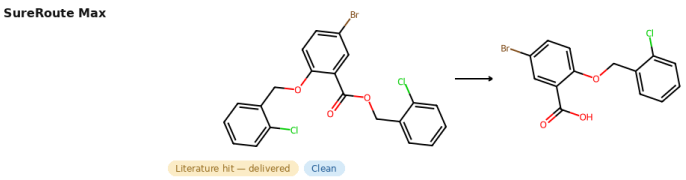
Method	Before → After
LocalRetro	19.7% → 0.0%
Chemformer	14.9% → 6.9%
Graph2Edits	15.4% → 0.6%
MEGAN	14.3% → 0.3%
MHNreact	15.7% → 0.0%
RetroKNN	14.9% → 0.0%
RootAligned	15.4% → 0.0%
DeepSeek V4	27.4% → 0.0%
GPT-5.5	27.4% → 0.9%
Claude Opus 4.8	18.9% → 0.3%
SureRoute Max	4.6% → 0.0%

Residual rate as a diagnostic. Reranking fully eliminates hallucination for candidate-rich models (LocalRetro, MEGAN) but leaves a small residual for Chemformer (6.9%)—because when a model's entire top-10 candidate pool consists of competition-flagged disconnections, there is no clean alternative for the reranker to promote. The residual rate after reranking is therefore itself informative: it measures whether a backbone's candidate *pool* contains a chemically sound option at all, independent of whether its own top-1 ranking found it.

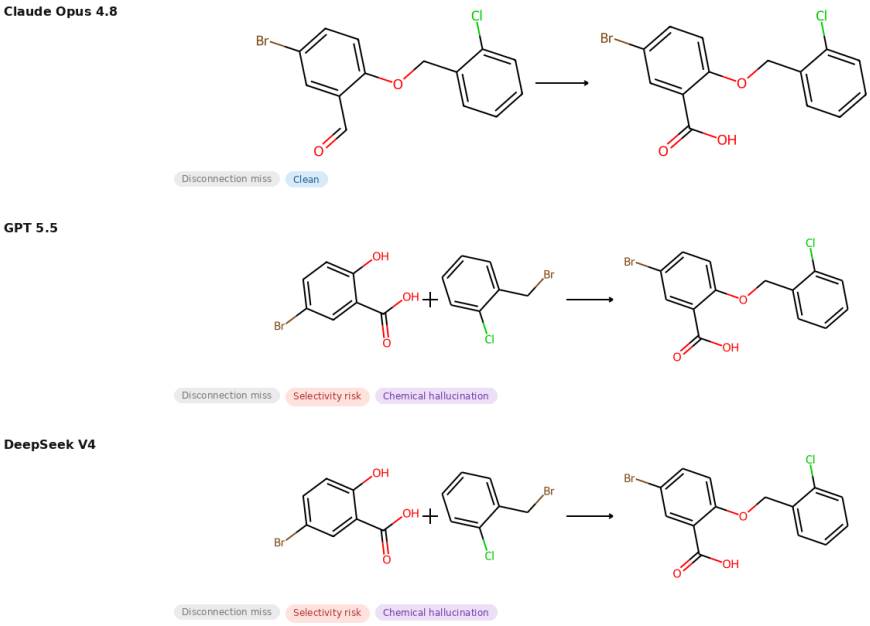
5.4 Two selectivity case studies

Sections 5.1–5.3 show *that* competing systems hallucinate; the cases below show *how*. We first examine two targets in detail, comparing every backbone's delivered top-1 route side by side, and then summarize two further cases. In each case the hallucination is not an artifact of one architecture: specialized models and frontier LLMs, trained and prompted independently, converge on the same chemically unexecutable disconnection.

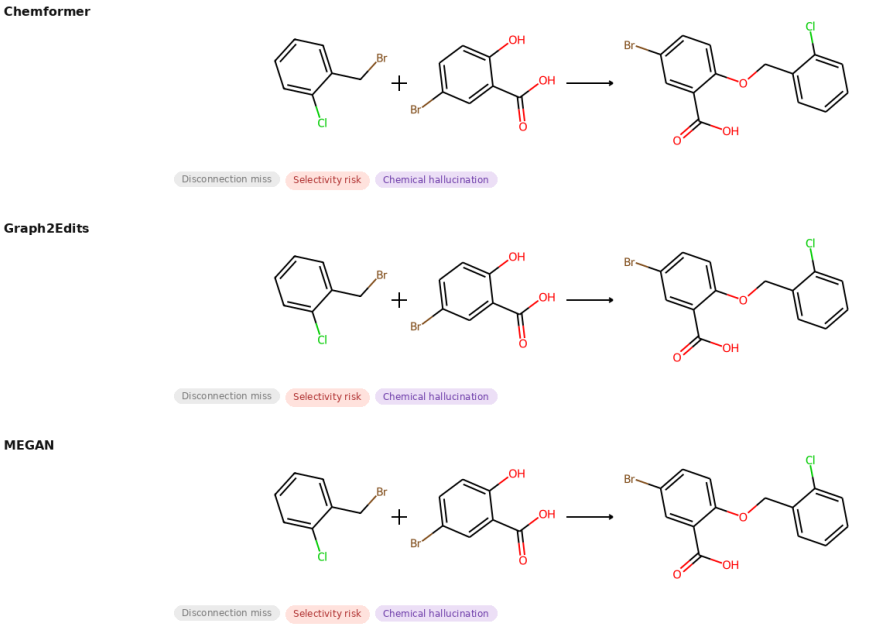
Case 1: chemoselectivity (carboxylate vs phenoxide) in a Williamson etherification



(a) SureRoute Max: literature ester-hydrolysis route.



(b) Frontier LLMs: Claude Opus 4.8 gives a clean non-reference oxidation; GPT-5.5 and DeepSeek V4 return the competitive ether disconnection.



(c) Chemformer, Graph2Edits, and MEGAN all return the same competitive Williamson disconnection.

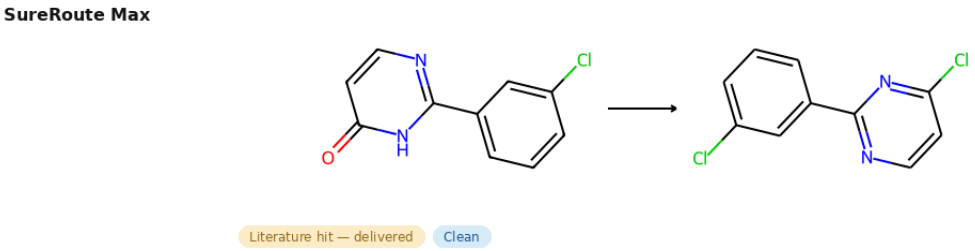
Figure 6: Case 1: chemoselectivity (carboxylate vs phenoxide). Specialized models and two frontier LLMs converge on a chemoselectively wrong Williamson disconnection; SureRoute retains the literature ester-hydrolysis route.

The target carries a free carboxylic acid *ortho* to a benzyl aryl ether. The obvious retrosynthetic move is to cut the ether bond, giving 5-bromosalicylic acid and 2-chlorobenzyl bromide. Atomically the disconnection balances, and it looks like a textbook Williamson ether synthesis. Chemically it does not work as written: the carboxylic acid is substantially more acidic than the phenol, so under the basic conditions the alkylation requires, the carboxylate is deprotonated preferentially and is alkylated first, delivering the benzyl *ester* rather than the desired aryl ether. Resolving this requires either protecting the acid or exploiting a large reactivity difference that the proposed conditions do not supply. The literature route avoids the ambiguity entirely by installing the ether earlier and unmasking the acid last, via a simple ester hydrolysis (Figure 6(a)).

Every specialized model we tested proposes the Williamson disconnection as its top-1 (Figure 6(c)), and two of the three frontier LLMs do the same (Figure 6(b)). ChemHarness flags all of them for O-alkylation competition. Claude Opus 4.8 is the one backbone that avoids the hallucination, proposing an aldehyde oxidation that ChemHarness marks clean — a chemically executable route, though not the reference one, and therefore still a recall miss under the protocol of Section 4.3. This is precisely the distinction the Chemical Hallucination metric is designed to expose: a disconnection miss and a chemical hallucination are different failures, and only the latter would be unsafe to reinforce in a self-improvement loop.

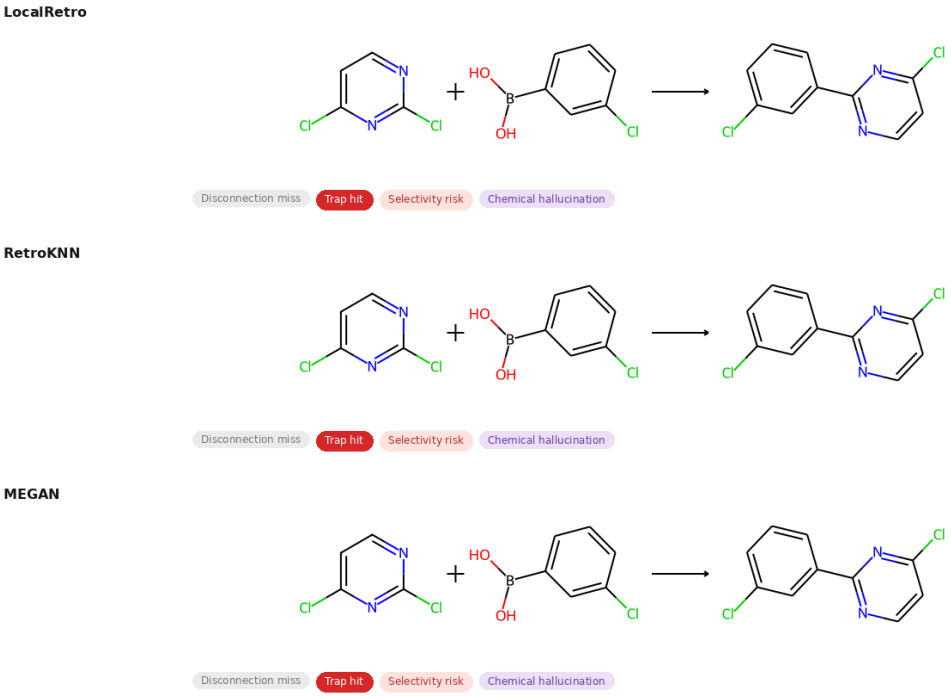
Case 2: regioselectivity — dichloropyrimidine (`ACYLSC046`)

The target contains a dichloropyrimidine scaffold with two potentially reactive C–Cl bonds. A natural retrosynthetic proposal couples 2,4-dichloropyrimidine with 3-chlorophenylboronic acid. However, the two chlorinated positions are not equally favored under conventional cross-coupling conditions: coupling is expected to occur preferentially at the more activated position, producing the undesired regioisomer rather than the target (Figure 7(a)). All seven specialized models place this disconnection at top-1, as does Claude Opus 4.8 (Figure 7(b)–(c)); GPT 5.5 gives an unmatched answer, and DeepSeek V4 does not return any route. ChemHarness flags each proposal for regioselectivity risk. The selectivity trap was independently annotated by an XtalPi chemist. Here, no backbone recovers the literature disconnection. Fluency and template coverage alike are blind to the selectivity problem, and only the verification layer separates the executable route from the hallucination one.

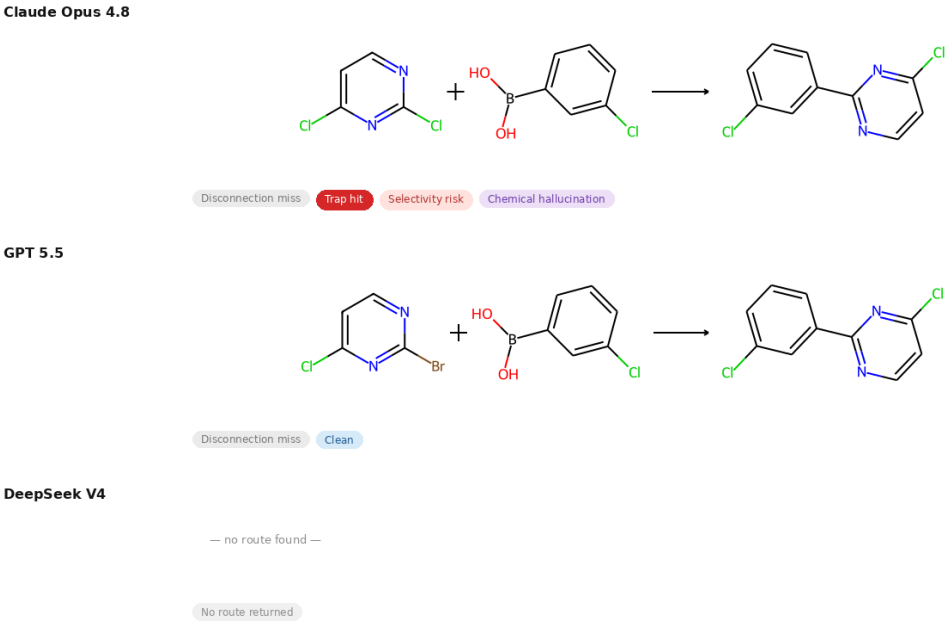


(a) SureRoute: matched with the literature route from the pyrimidinone.

Figure 7: Case 2: regioselectivity in a dichloropyrimidine coupling. Continued on the next page.



(b) Specialized retrosynthesis models such as LocalRetro, RetroKNN, and MEGAN all return the same selectivity-unsafe disconnection.

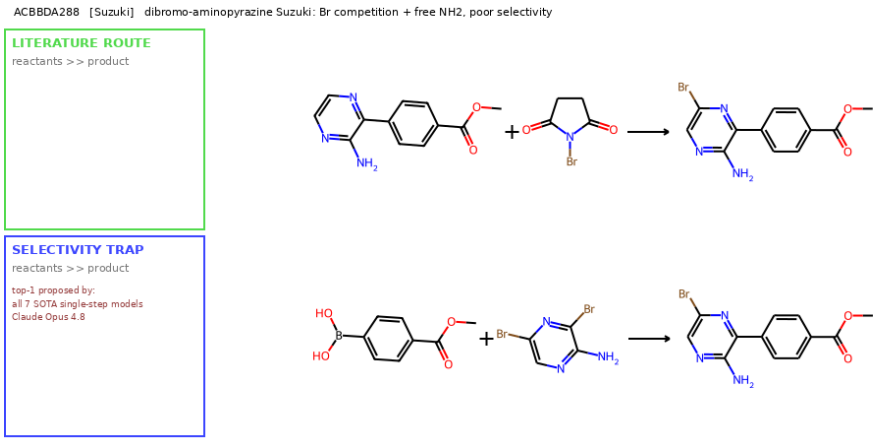


(c) Claude Opus 4.8 proposes the same disconnection without resolving the competing C–Cl site.

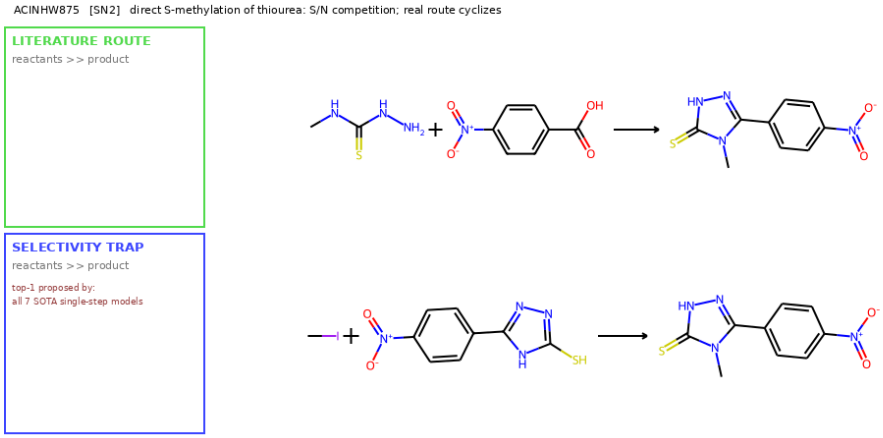
Figure 7: Case 2: regioselectivity in a dichloropyrimidine coupling. Competing C–Cl sites make the coupling position ambiguous; SureRoute ranks the literature-supported pyrimidinone route first.

Further Chemical Hallucination cases

Additional cases follow the same pattern — all seven single-step models (and, in most cases, one or more LLMs) converge on a single flagged disconnection:



(a)



(b)

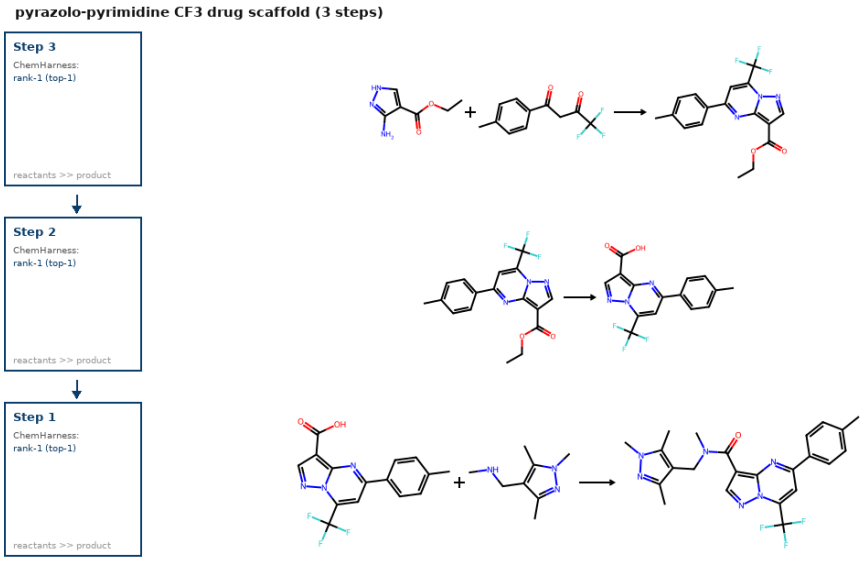
Figure 8: Further selectivity-hallucination cases. Blue boxes show literature routes; grey boxes show atom-mapping consistent but selectivity-unsafe top-1 predictions. ChemHarness screens precisely this failure mode.

- **Dibromo-aminopyrazine Suzuki coupling** (`ACBBDA288`): competing bromides plus an unprotected free amine create a selectivity problem; the literature route disconnects at a single, unambiguous aromatic ring position.
- **Thiourea *S*-methylation** (`ACINHW875`): competing *S*- vs. *N*-methylation leaves the product structure ambiguous; the literature route proceeds through a cyclization step that fixes the reactive site.

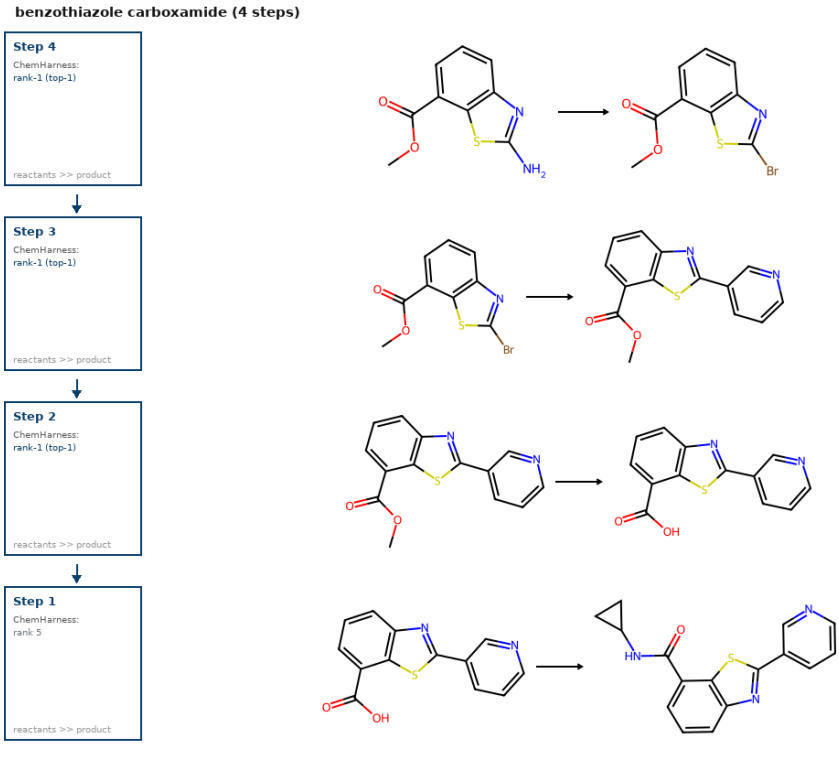
In each, all seven single-step models' top-1 prediction lands on a chemically hallucination route; only SureRoute's ChemHarness-based reranking anchors the top-1 slot on the literature route (Figure 8).

5.5 Multi-step extension

While the main evaluation is single-step, we additionally verified that the same reliability advantage compounds over multi-step routes, using a separate set of literature targets (3–5 steps) with fully documented experimental routes. Under step-wise interactive prediction (ChemHarness re-evaluated at each step), we recovered the complete literature route—every step, in the correct order, with the correct disconnection ranked first at nearly every individual step (Figure 9). Beyond these benchmarks, hundreds of expert chemists at XtalPi interact with SureRoute in production, annotating mis-ranked routes and continually evolving the system.



(a)



(b)

Figure 9: Two literature multi-step routes recovered by ChemHarness-guided step-wise prediction. Blue annotations report the rank of the reference disconnection at each step; the correct choice remains top-ranked at nearly every step.

6 Conclusion

We introduced SureRoute, a chemically verified self-improving retrosynthesis platform that pairs a multi-model disconnection ensemble and literature retrieval with ChemHarness, a rule-based, mechanism-aware reliability layer that explicitly screens for functional-group competition and

mechanistic implausibility. On an internal 350-molecule benchmark, SureRoute achieves 2.2–3.5$\times$ the top-1 recall of both SOTA single-step models and frontier LLMs, and—more importantly—a Chemical Hallucination rate of 4.6%, a 4–6$\times$ reduction relative to frontier LLMs. We showed that this reliability gain is not tied to our own ensemble: applying ChemHarness reranking to any backbone's raw candidates, including LLMs it was never tuned against, drives ChemHarness-detectable top-1 Chemical Hallucination toward near-zero, establishing ChemHarness as a model-agnostic, pluggable reliability layer rather than a component specific to one system.

More broadly, we position SureRoute's contribution within a wider trend of self-improving retrosynthesis and scientific-AI systems, where the central open problem has shifted from generation capability to verification capability: a self-improvement loop is only as trustworthy as whatever decides which self-generated outputs to reinforce, and a loop that grades its own output risks compounding its own mistakes rather than correcting them. ChemHarness is designed to occupy the space between wet-lab-grade trust and unit-test-grade cost—an executable, auditable verifier that lets SureRoute close a rule-refinement loop over real production failures without either the expense of physical validation or the risk of self-graded reward. This verifier-anchored, compounding advantage is, we argue, difficult for a general-purpose LLM to replicate through scale or prompting alone, because it depends on continually accumulating and correcting against real deployment failures rather than on any fixed amount of pretraining.

References

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Advances in Neural Information Processing Systems*, volume 30, pages 5360–5370, 2017.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624:570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Benjamin Burger, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M. Alston, Buyi Li, Rob Clowes, Nicola Rankin, Brandon Harris, Reiner Sebastian Sprick, and Andrew I. Cooper. A mobile robotic chemist. *Nature*, 583:237–241, 2020. doi: 10.1038/s41586-020-2442-2.

Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: Learning retrosynthetic planning with neural guided A* search. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1608–1616. PMLR, 2020.

Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021. doi: 10.1021/jacsau.1c00246.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009. doi: 10.1145/1571941.1572114.

Hanjun Dai, Chengtao Li, Connor W. Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, volume 32, pages 8870–8880, 2019.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, Hongru Wang, Han Xiao, Yuhang Zhou, Shaokun Zhang, Jiayi Zhang, Jinyu Xiang, Yixiong Fang, Qiwen Zhao, Dongrui Liu, Qihan Ren, Cheng Qian, Zhenhailong Wang, Minda Hu, Huazheng Wang, Qingyun Wu, Heng Ji, and Mengdi Wang. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=CTr3bovS5F.

Jiasheng Guo, Chenning Yu, Kenan Li, Yijian Zhang, Guoqiang Wang, Shuhua Li, and Hao Dong. Retrosynthesis zero: Self-improving global synthesis planning using reinforcement learning. *Journal of Chemical Theory and Computation*, 20(11):4921–4938, 2024. doi: 10.1021/acs.jctc.4c00071.

Siqi Hong, Hankz Hankui Zhuo, Kebing Jin, Guang Shao, and Zhanwen Zhou. Retrosynthetic planning with experience-guided monte carlo tree search. *Communications Chemistry*, 6:120, 2023. doi: 10.1038/s42004-023-00911-8.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55, 2025. doi: 10.1145/3703155.

Ilia Igashov, Arne Schneuing, Marwin H. S. Segler, Michael M. Bronstein, and Bruno Correia. RetroBridge: Modeling retrosynthesis with markov bridges. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=770DetV8He.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38, 2023. doi: 10.1145/3571730.

Pavel Karpov, Guillaume Godin, and Igor V. Tetko. A transformer model for retrosynthesis. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, volume 11731 of *Lecture Notes in Computer Science*, pages 817–830. Springer, 2019. doi: 10.1007/978-3-030-30493-5_78.

Junsu Kim, Sungsoo Ahn, Hankook Lee, and Jinwoo Shin. Self-improved retrosynthetic planning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5486–5495. PMLR, 2021.

Heeseung Lee, Hyuk Jun Yoo, Hye Su Jang, Byeongho Park, Yang Jeong Park, and Sang Soo Han. Toward self-driving laboratory 2.0 for chemistry and materials discovery. *Materials Horizons*, 13: 4712–4739, 2026. doi: 10.1039/D5MH01984B.

Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3(10):1103–1113, 2017. doi: 10.1021/acscentsci.7b00303.

Guoqing Liu, Di Xue, Shufang Xie, Yingce Xia, Austin Tripp, Krzysztof Maziarz, Marwin H. S. Segler, Tao Qin, Zongzhang Zhang, and Tie-Yan Liu. Retrosynthetic planning with dual value networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22266–22276. PMLR, 2023.

Wei Liu, Jiangtao Feng, Hongli Yu, Yuxuan Song, Yuqiang Li, Shufei Zhang, Lei Bai, Wei-Ying Ma, and Hao Zhou. Retro-R1: LLM-based agentic retrosynthesis. In *Advances in Neural Information Processing Systems*, volume 38, pages 70709–70737, 2026.

Daniel Mark Lowe. *Extraction of Chemical Structures and Reactions from the Literature*. PhD thesis, University of Cambridge, 2012.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594, 2023.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin H. S. Segler. Re-evaluating retrosynthesis algorithms with Syntheseus. *Faraday Discussions*, 256:568–586, 2025. doi: 10.1039/D4FD00093E.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

RDKit contributors. RDKit: Open-source cheminformatics. https://www.rdkit.org, 2024. Software; cite the version-specific Zenodo DOI when reproducibility requires an exact release.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Marwin H. S. Segler and Mark P. Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry—A European Journal*, 23(25):5966–5971, 2017. doi: 10.1002/chem.201605499.

Marwin H. S. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555:604–610, 2018. doi: 10.1038/nature25978.

Shuai Shao, Qihan Ren, Dongrui Liu, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, and Jing Shao. Your agent may misevolve: Emergent risks in self-evolving LLM agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lS1gWUHbfx.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017. doi: 10.1038/nature24270.

Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature*, 624:86–91, 2023. doi: 10.1038/s41586-023-06734-w.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, Arman Cohan, and Mark Gerstein. ChemAgent: Self-updating memories in large language models improves chemical reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kuhIqeVg0e.

Igor V. Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11: 5575, 2020. doi: 10.1038/s41467-020-19266-y.

Austin Tripp, Krzysztof Maziarz, Sarah Lewis, Marwin H. S. Segler, and José Miguel Hernández-Lobato. Retro-Fallback: Retrosynthetic planning in an uncertain world. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=dl0u4ODCuW.

Shufang Xie, Rui Yan, Peng Han, Yingce Xia, Lijun Wu, Chenjuan Guo, Bin Yang, and Tao Qin. RetroGraph: Retrosynthetic planning with graph search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2120–2129, 2022. doi: 10.1145/3534678.3539446.

Robin Yadav, Qi Yan, Guy Wolf, Joey Bose, and Renjie Liao. RetroSynFlow: Discrete flow matching for accurate and diverse single-step retrosynthesis. In *Advances in Neural Information Processing Systems*, volume 38, pages 55229–55258, 2026.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. RetroXpert: Decompose retrosynthesis prediction like a chemist. In *Advances in Neural Information Processing Systems*, volume 33, pages 11248–11258, 2020.

Xunjian Yin, Xinyi Wang, Liangming Pan, Li Lin, Xiaojun Wan, and William Yang Wang. Gödel agent: A self-referential agent framework for recursively self-improvement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27890–27913, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1354.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 15476–15488, 2022.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025a.

Situo Zhang, Hanqi Li, Lu Chen, Zihan Zhao, Xuanze Lin, Zichen Zhu, Bo Chen, Xin Chen, and Kai Yu. Reasoning-driven retrosynthesis prediction with large language models via reinforcement learning. *arXiv preprint arXiv:2507.17448*, 2025b.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, 2025c. doi: 10.1162/coli.a.16.

Chenyang Zuo, Siqi Fan, Yizhen Luo, and Zaiqing Nie. Reinforced reasoning for end-to-end retrosynthetic planning. *arXiv preprint arXiv:2603.29723*, 2026.

A LLM Evaluation Protocol

We evaluate frontier large language models for retrosynthesis using a unified prompt template. All generated candidates are processed using the same downstream standardization and matching pipeline as specialized retrosynthesis models (Section 4.3). The full prompt consists of a system instruction and a user message containing two in-context demonstrations followed by the query.

A.1 Prompt Template

System

```
You are an expert organic chemist specializing in single-step retrosynthesis.
Given a target molecule, propose reactant sets that could be combined in one
reaction step to synthesize it.
```

User

Demonstration 1 (amide coupling)

```
Target:
O=C(Nc1ccccc1)c1ccccc1
Answer:
O=C(O)c1ccccc1.Nc1ccccc1
O=C(Cl)c1ccccc1.Nc1ccccc1
```

Demonstration 2 (Suzuki coupling)

```
Target:
CCOC(=O)c1ccc(-c2ccccc2)cc1
Answer:
CCOC(=O)c1ccc(Br)cc1.OB(O)c1ccccc1
CCOC(=O)c1ccc(Br)cc1.CC1(C)OB(c2ccccc2)OC1(C)C
```

Query

```
Propose up to 5 distinct single-step retrosynthetic disconnections for the
target below.

Rules:
- Each line represents one reactant set.
- Reactants are separated by ".".
- Order candidates from most to least likely.
- Output only reactant SMILES.
- Do not provide explanations or reaction arrows.
- Reactants should represent purchasable-style building blocks.

Target:
{target SMILES}
Answer:
```

This prompt elicits ranked candidate lists rather than a single prediction, enabling direct comparison between LLMs and specialized retrosynthesis models under the same recall@K evaluation protocol.

B Target Visualization

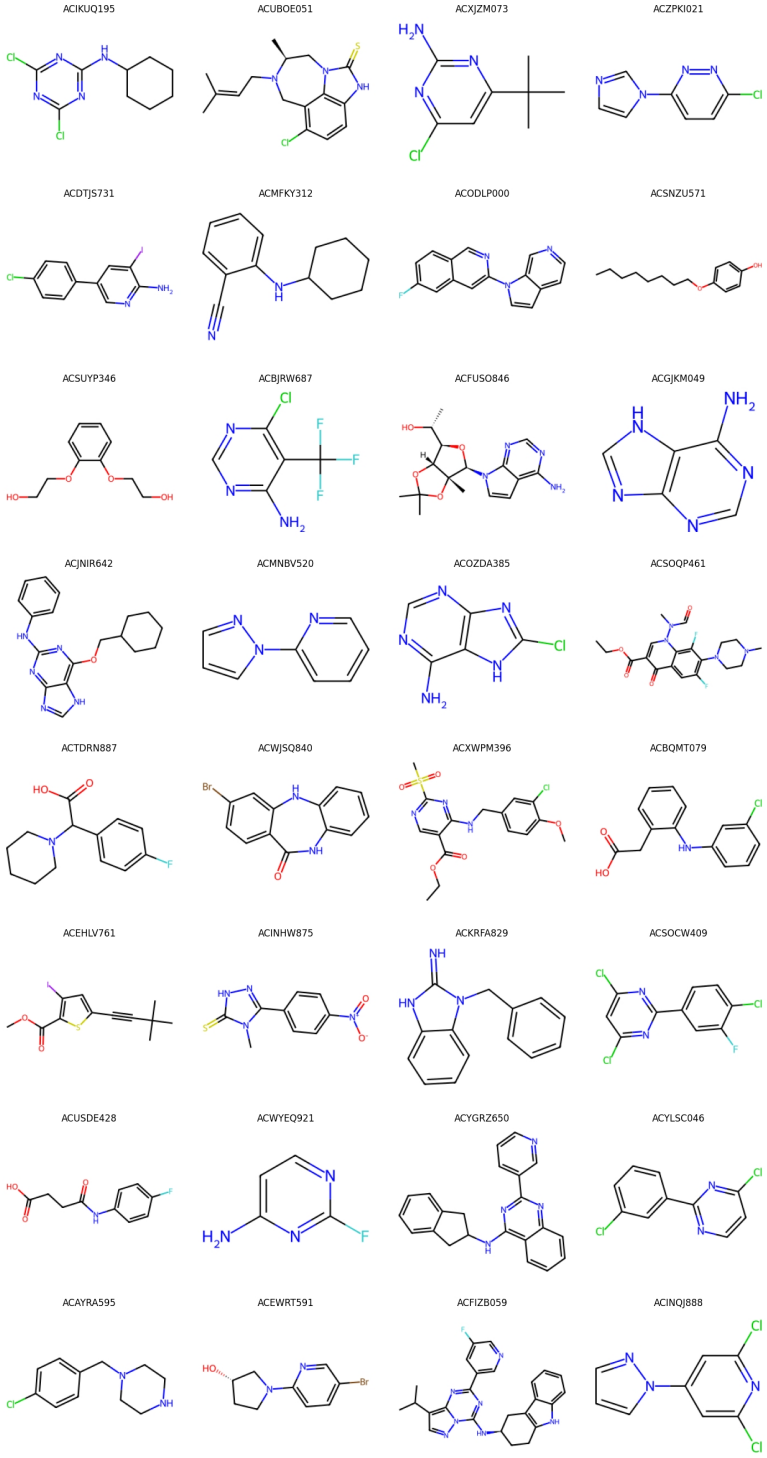


Figure 10: Target visualization of the industrial retrosynthesis evaluation set.