Trustworthy Retrosynthesis: Eliminating Hallucinations with a Diverse Ensemble of Reaction Scorers

Michał Sadowski Molecule.one
Tadija Radusinović Molecule.one
Maria Wyrzykowska Molecule.one
Lukasz Sztukiewicz Molecule.one
Jan Rzymkowski Molecule.one
Piotr Kozakowski Molecule.one
Mikołaj Sacha Molecule.one
Ruard van Workum Molecule.one
Stanislaw Kamil Jastrzebski Molecule.one

Abstract

Retrosynthesis is one of the domains transformed by the rise of generative models, and it is one where the problem of nonsensical or erroneous outputs (hallucinations) is particularly insidious: reliable assessment of synthetic plans is time-consuming, with automatic methods lacking. In this work, we present RetroTrim, a retrosynthesis system that successfully avoids nonsensical plans on a set of challenging drug-like targets. Compared to common baselines in the field, our system is not only the sole method that succeeds in filtering out hallucinated reactions, but it also results in the highest number of high-quality paths overall. The key insight behind RetroTrim is the combination of diverse reaction scoring strategies, based on machine learning models and existing chemical databases. We show that our scoring strategies capture different classes of hallucinations by analyzing them on a dataset of labeled retrosynthetic intermediates. To measure the performance of retrosynthesis systems, we propose a novel evaluation protocol for reactions and synthetic paths based on a structured review by expert chemists. Using this protocol, we compare systems on a set of 32 novel targets, curated to reflect recent trends in drug structures. While the insights behind our methodology are broadly applicable to retrosynthesis, our focus is on targets in the drug-like domain. By releasing our benchmark targets and the details of our evaluation protocol, we hope to inspire further research into reliable retrosynthesis.

1 Introduction

The advent of deep generative modeling has transformed a broad range of domains, and resulted in impressive applications such as photorealistic image synthesis, code generation, and automated theorem proving [Rombach et al., 2022, Li et al., 2022, Yang et al., 2023]. The undisputed efficacy of generative models is nonetheless hindered by the possibility of factually wrong or outright nonsensical output, often called "hallucinations" [Sahoo et al., 2024]. In some applications, such as automated theorem proving, hallucinations can be curtailed with formal verification of output. In others, they can only be mitigated through partial and often imprecise verification.

Retrosynthesis, the task of constructing synthetic routes — sequences of chemical reactions that lead to a desired target molecule from simpler precursors — is another domain that has undergone significant developments with the rise of generative modeling [Coley et al., 2018]. Modern systems

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

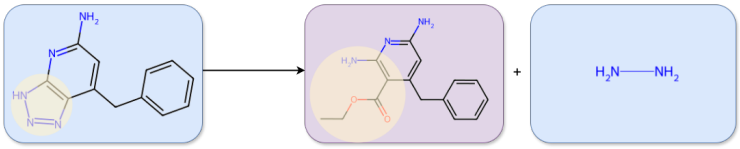


Figure 1: An example of grossly incorrect (hallucinated) reaction generated by a Single-Step Retrosynthesis model. A PhD-level chemist recognizes that the only reasonable atom mapping between the substrates and the product is one where the reaction center is an ortho-amino benzoate converting into a triazole (highlighted in yellow). It does not belong to any commonly known reaction class, and further investigation involving extensive searches of synthetic databases yields no examples that would inform what reagents and conditions could induce such a reaction.

typically generate complete synthetic routes by iteratively applying predictions from a single-step retrosynthesis (SSR) model within a graph-based search algorithm. When reactions have no known chemical precedent or display features such as unstable reagents, they jeopardize the validity of the synthetic path they are part of. Generative SSR models inherit the limitations of generative modeling: the reactions they propose are frequently hallucinated, that is, chemically unsound [Torren-Peraire et al., 2024]. An example of a hallucinated reaction is shown in Figure 1.

In this work, we propose RetroTrim, a retrosynthesis system that pairs a performant SSR generator with an ensemble of complementary reaction scorers as a plausibility filter. Among common retrosynthesis systems, RetroTrim is the only one that avoids *all* hallucinated reactions in proposed synthetic plans for unpublished challenging drug-like targets, while also finding flawless routes for the largest number of targets.

Our approach in developing RetroTrim is based on the principle of ensemble learning [Hansen and Salamon, 1990]: combining diverse scorers with distinct error patterns yields a more robust and accurate assessment of reaction plausibility. The ensemble scorer (MetaScorer) functions as an external validator of the generative SSR model. It prunes all reactions below a given score threshold from the synthesis tree, curtailing expansion of hallucinated nodes. For an overview see Fig. 2. MetaScorer combines three complementary filters, detailed in Secs. 3.1–3.3. Reaction Prior (RP) is a Transformer-based learned plausibility model [Vaswani et al., 2017] that mimics broad chemist judgments about reaction feasibility. Reaction Graph Plausibility (RGP) is a graph classifier trained against synthetic negatives, with forward negatives targeting selectivity errors and retrosynthetic negatives capturing broader functional-group incompatibilities. Reaction Retrieval Score (RRS) checks for experimental precedent and is most useful for gross hallucinations, such as reactions with no close references or large reactant-product mismatches.

We compare performances of RP, RGP and RRS on a dataset of retrosynthetic intermediates. We show that each scorer excels at filtering different kinds of hallucinations, and that the MetaScorer further improves performance in most areas.

In identifying hallucinations of retrosynthesis systems, common automated metrics can at best serve as surrogates: algorithmic procedures cannot model chemical plausibility, and exact literature matches are rare for complex targets. In practice, a synthetic route must be judged by expert chemists, who can reject reactions lacking known synthetic methodology without experimental verification.

For this reason, we evaluate retrosynthesis systems and reaction scorers with a structured validation protocol in which PhD-level chemists assign each reaction an issue type and severity level. The protocol distinguishes hallucinated reactions from milder plausibility concerns and assigns each pathway the severity of its weakest reaction; details are provided in Sec. 4 and Appendix F.

All retrosynthesis systems are evaluated on thirty-two challenging drug-like targets (Appendix G) that match common features of modern drugs while avoiding close literature analogs. This prevents train-test leakage while remaining representative of medicinal chemistry.

Our main contributions are as follows:

- We propose RetroTrim, which is the only evaluated method to reject all hallucinated reactions on challenging targets while finding the most flawless synthetic routes.
- We show that RetroTrim's three reaction filters have complementary strengths across issue types, and that combining them yields the best system.
- We introduce a chemist-led reaction labeling protocol covering seven issue types and four severity levels for granular evaluation of reactions and routes.
- We release 32 unpublished drug-like targets as a challenging retrosynthesis test set.
- We release a benchmark of 500 expert-labeled reactions, together with the code of our reaction scorers, the corresponding models trained on USPTO, and their integration with the Syntheseus search framework [Maziarz et al., 2025], at `https://github.com/molecule-one/retrotrim`.

2 Related work

To contextualize our proposed method, this section examines the workflow in automated retrosynthesis, covering both the generation of reactions by Single-Step Retrosynthesis (SSR) models and techniques used for their validation.

2.1 Multi-step retrosynthesis

Multi-step retrosynthesis constructs complete synthetic routes by iteratively applying single-step retrosynthesis (SSR) predictions. Classical search strategies, such as Monte Carlo tree search [Segler et al., 2018] and A* search (Retro*) [Chen et al., 2020], use SSR models to expand compounds during generation of synthetic pathways.

SSR models span template-based methods [Coley et al., 2017, Segler and Waller, 2017, Dai et al., 2019a], template-free neural approaches, including seq2seq, Transformer, and GNN-based models [Liu et al., 2017, Karpov et al., 2019, Zhong et al., 2022, Sacha et al., 2021], and semi-template-based systems [Somnath et al., 2021, Yan et al., 2020]; all categories remain prone to erroneous output [Coley et al., 2018].

Evaluating single-step retrosynthesis (SSR) models remains challenging. Traditional metrics, such as top-k accuracy, measure whether the correct reactants appear among the top predicted candidates, but they provide limited insight into chemical plausibility and alternative valid reactions.

To address this, round-trip accuracy [Schwaller et al., 2020] passes predicted reactants through a forward reaction model and checks whether the original product is recovered. It is a better proxy for chemical validity than top-k accuracy, but its relationship to actual plausibility remains unclear.

Multi-step route evaluation is typically limited to route-level properties such as step count, branching factor, or synthesizability scores [Maziarz et al., 2025], which provide little insight into chemical correctness. Human evaluation should address this gap, but the two main examples remain limited in scope or reproducibility. Segler et al. [2018] used a double-blind A/B test to compare literature and model-generated routes, yet measured preference rather than system-error rates and covered only 9 targets. GRASP [Yu et al., 2022] also uses human assessment, but does not release a per-reaction annotation protocol or disentangle feasibility from efficiency and simplicity in its composite score. In contrast, we conduct fine-grained expert error analysis of routes from 4 modern retrosynthesis systems and 6 RetroTrim variants, including full RetroTrim, on 32 challenging targets, and release the protocol, annotated examples, and per-reaction labeling effort.

2.2 Mid-search reaction validation

A common way to improve pathway validity is to use a plausibility model as a filter. Such models are typically classifiers trained to distinguish correct reactions from artificial negatives, often created by perturbing positives, for example by randomly modifying substrates or swapping the product with a chemically similar one [Segler et al., 2018, Genheden et al., 2020a]. IBM RXN uses a forward model trained only on positive data in the same role [Schwaller et al., 2020].

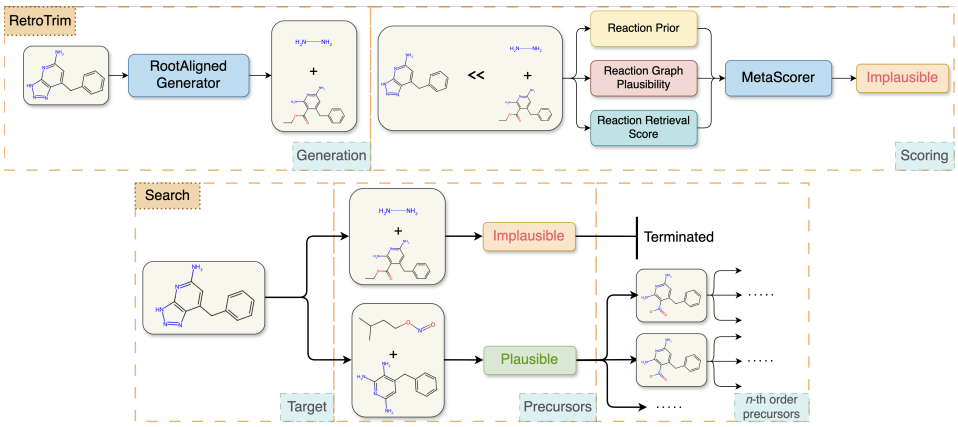


Figure 2: Visualization of **RetroTrim** (above) and **retrosynthetic search with plausibility filtering** (below). RetroTrim encompasses a generator, which proposes precursor molecules for a given target, and a scorer, which evaluates the plausibility of the generated reaction. In the search process, plausible precursors are expanded further, until we arrive at commercially available starting materials. Implausible reactions terminate the search branch. The search concludes when a complete synthetic route from commercially-available starting materials to the target molecule is found.

Another strategy scores generated reactions by likelihood during search. This can use template-free models, such as Molecular Transformer confidence from output token probabilities [Schwaller et al., 2019], or template-based softmax scores, as in RetroFallback [Tripp et al., 2023]. RetroGFN [Gaiński et al., 2024] improves predictions using feedback from a forward model and round-trip accuracy, while RetroChimera [Maziarz et al., 2024] selects plausible predictions from an ensemble of diverse generators.

Finally, retrieval-based validation grounds predictions in chemical precedent, mirroring a chemist's literature search. Retrieval-augmented methods such as RARB retrieve similar molecules to guide reactant generation [Qiao et al., 2025]. In this context, RetroTrim is the first system to use retrieved reactions to filter predictions from a reaction generator.

Large language models have recently been applied to both roles. Bran et al. [2025] use LLM-assigned strategy and feasibility judgments to steer retrosynthetic search and to assess routes from retrosynthesis systems and the literature, but without human verification of the resulting reactions and without addressing hallucinations. Hassen et al. [2025], concurrent with our work, also involve expert chemists, but only to check individual LLM-predicted steps, not to score reactions for plausibility inside a search.

To date, none of these approaches have been extensively validated through human expert evaluation.

3 Methods

RetroTrim combines an SSR reaction generator with a reaction scorer. Its key innovation is the scorer, built around a central Reaction Prior model and supported by Reaction Graph Plausibility and Reference Reaction Scorer components that complement it on specific error types. Together, they form the MetaScorer, which provides the final assessment of reaction correctness.

3.1 Reaction prior

The Reaction Prior (RP) is a novel method inspired by how experienced chemist reason about reactions: considering the reaction globally, assessing the reaction center, and comparing it to alternative reactions that could occur. The RP score S_{final} is thus a weighted product of three components, each first normalized to $(0, 1)$: $S_{final} = \tilde{S}_{RP}^{\alpha}\, \tilde{S}_{Regio}^{\beta}\, \tilde{S}_{RC}^{\gamma}$. Here, S_{RP} is the Reaction Prior (global) Score, S_{RC} is the Reaction Center Score, and S_{Regio} is the Regioselectivity Score, with α, β, and γ serving as weighting factors. These weights can be tuned to optimize a metric of interest. In practice, we find $\alpha = 1$, $\beta = 2.5$ and $\gamma = 1.5$ to be reasonable default values for drug-like targets.

Appendix C.1 describes how they were selected. RP is implemented as a decoder-side autoregressive scorer based on BART [Lewis et al., 2019a], where substrates and products are processed by the decoder, trained for next-token prediction by minimizing cross-entropy loss.

Reaction prior score (S_{RP}) This score reflects the overall feasibility of the reaction. It is proportional to the log probability assigned by the model to the reaction sequence. Analogously to language models, which assign higher probabilities to sequences close to the training set, this score acts as a measure of reaction similarity to the dataset on which the model is trained on. In this respect, it mimics the way in which a chemist would look for precedent for the overall transformation. We normalize the log probability by the square root of the number of characters (L) in the reaction string, which we find to maximize predictive performance in practice: $S_{RP} = \frac{1}{\sqrt{L}} \log P(\text{reaction})$.

Reaction center score (S_{RC}) When evaluating reactions, chemists pay special attention to the sites where changes in connectivity occur. The S_{RC} score analogously evaluates the model's confidence in the identified reactive sites. It is proportional to the sum of log probabilities of tokens representing atoms in the reaction's center: $S_{RC} = \sum_{i \in \text{reaction center}} \log P(\text{token}_i)$.

Regioselectivity score (S_{Regio}) This component quantifies reaction site specificity. It is calculated by comparing the reaction at the given site with the alternative sites where the reaction might occur, aggregated into a soft count N of competing sites: $S_{Regio} = \frac{2}{1+N}$, so that the score is highest when no competing site is found. Details are provided in Appendix C.2. This score component reflects the tendency for chemists to evaluate whether the particular site where a reaction occurs is preferred compared to potential alternatives.

3.2 Reaction graph plausibility

A Graph Attention Network (GAT) [Veličković et al., 2017] is trained to differentiate chemically valid reactions from implausible ones. Training uses reaction datasets for positive examples and synthetic negative examples generated through forward and two-step backward template applications. This approach is similar to those proposed by [Segler et al., 2018, Genheden et al., 2020a], but it uses a graph neural network instead of a feedforward network with fingerprint inputs, and employs more sophisticated artificial negative reactions, generated not only by applying random templates in the forward direction but also through retro-synthetic random template applications, which increases the diversity of types of generated incorrect reactions. Details of GAT featurization are described in Appendix D.

3.3 Reference reaction scorer

We developed a structured retrieval pipeline that extracts chemical precedent information through a two-tiered lookup over canonical transformation keys based on bond change patterns. First, *coarse-grained indexing* extracts connected components of the reaction center and applies atom mapping (with RXNMapper [Schwaller et al., 2021]) to identify the underlying transformation pattern. Reactions share a coarse key if their transformation patterns are identical. Then, *fine-grained indexing* extends the coarse-grained approach by incorporating chemically significant substructures — aromatic systems and conjugated double bonds - into the key.

Our Reference Reaction Retrieval Scorer (RRS) quantifies reaction plausibility through a logarithmic transformation of the reference reaction count obtained from the index. At query time we first look up the fine-grained bucket of the candidate reaction; if it contains at least one reference, its size is returned as the reference count, otherwise we fall back to the coarse-grained bucket size:

$$p(\texttt{reaction}) = \log(n_{\texttt{ref}}(\texttt{reaction}) + 1) \tag{1}$$

where $n_{\texttt{ref}}(\texttt{reaction})$ is the size of the matched bucket (fine-grained if available, coarse-grained otherwise). Full implementation details are provided in Appendix E.

3.4 MetaScorer aggregation

To improve reaction filtering, our MetaScorer integrates scores from Reaction Prior ($s_{\texttt{RP}}$), Reaction Graph Plausibility ($s_{\texttt{RGP}}$), and empirical precedents ($n_{\texttt{ref}} > 0$) retrieved via the pipeline in Sec. 3.3.

This hybrid approach mitigates the weaknesses of purely data-driven or precedent-based methods. Both component scores are first calibrated on a held-out set of chemist-annotated reactions, disjoint from the one used for evaluation; we write $\tilde{s}$ for a calibrated score. The continuous score, which we use to rank reactions, is described by equation $s_{\texttt{META}} = \max(\tilde{s}_{\texttt{RGP}}, \tilde{s}_{\texttt{RP}})$ if $n_{\texttt{ref}} > 0$ (0 otherwise). The MetaScorer operates on individual reactions and does not depend on the generator, so it can also be applied as a reaction-level filter to candidates produced by other systems.

For binary classification tasks and search, reactions are filtered by a separate decision rule that applies a predefined threshold to each scorer, which can be selected through grid search to balance precision and recall:

$$d_{\texttt{META}} = \begin{cases} 1 & \text{if } s_{\texttt{RGP}} > \texttt{thr}_{\texttt{RGP}} \text{ and } s_{\texttt{RP}} > \texttt{thr}_{\texttt{RP}} \text{ and } n_{\texttt{ref}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

3.5 Generator and integration with search (Retro*)

For the generator part of RetroTrim, we use the encoder-decoder BART architecture [Lewis et al., 2019b] trained on root-aligned SMILES [Zhong et al., 2022], where the product (target) is processed by the encoder, and the substrates are generated by the decoder. We call this generator RootAligned. At each expansion, candidate precursor sets are decoded with deterministic diverse beam search; the decoding settings are given in Appendix K. The calibrated MetaScorer is used during multi-step retrosynthesis search to improve the quality of the pathways predicted by the BART generator. We integrate the scorer into the Retro* [Chen et al., 2020] search framework as a reaction filtering mechanism. Reactions are pruned from the search tree if $d_{\texttt{META}}$ defined in 2 is equal to 0.

4 Human evaluation

We curated a dataset of over 4,500 reactions generated by our SSR models. Each reaction was evaluated and labeled by PhD-level chemists into expert-defined categories, creating the first comprehensive dataset of its kind. This dataset lets us evaluate error patterns in reaction plausibility scorers. A subset of 500 reactions is released as a community benchmark.

Reaction evaluation protocol. We designed the protocol to systematically evaluate predicted reactions based on expert-defined heuristics. Reactions were rated using a four-point confidence scale: *Nonsense*, *Rather not*, *Worthwhile*, and *Safe bet*. *Safe bet* reactions are considered fully reliable; *Worthwhile* reactions remain plausible but carry a moderate risk of failure; *Rather not* reactions are associated with a high risk of major difficulties; and *Nonsense* reactions are effectively infeasible, i.e. hallucinated. For the system to be reliable, valid pathways should consist primarily of *Safe bets* reactions. The presence of *Nonsense* reactions effectively invalidates a pathway, while the presence of a *Rather Not* reaction may still be acceptable in target-oriented synthesis when no alternatives exist. Reactions that are not a *Safe bet* receive an additional label specifying the cause of their incorrectness, chosen from: *Reactants mismatch*, *Unstable*, *Magic*, *One pot*, *Reactivity*, *Functional group incompatibility*, and *Selectivity*. These error categories correlate with confidence levels to varying degrees: for example, *Magic* errors almost always map to *Nonsense*, while *Selectivity* issues more often correspond to *Worthwhile* or *Rather Not*. Otherwise, a reaction is assigned a *No Problem* label. A detailed description of the evaluation framework is provided in Appendix F.

Annotation effort. Over 4,500 reactions were annotated by twelve PhD-level synthetic chemists in total. Further details of the annotation team are provided in Appendix F. Each reaction took 3–6 minutes on average, corresponding to over 300 person-hours across the full dataset. At path level, this implies about 8–16 hours of PhD-level chemist review per system, since each system produces roughly 160 reactions across the 32 targets. Annotation was blind: the source system was anonymized, and reactions were randomly assigned across chemists. To assess consistency, approximately 20% of reactions were double-annotated by two independent chemists, with 88.6% agreement on the four-point confidence scale, 96.6% agreement within one scale point, and 96.5% agreement on the binary evaluation label. About 48% of disagreements were *Worthwhile–Safe bet* distinctions and therefore did not change the binary label; the remainder concerned mainly the amount of literature precedent required, the severity of selectivity or reactivity liabilities, and whether telescoped transformations count as a single step. These cases were resolved through discussion and literature review, and the resulting clarifications were added to the annotation guidelines.

Interpretation of the scale. The four-tier scale should be read as a practical utility scale for synthetic chemists rather than as absolute correctness labels. Laboratory execution is the gold standard for validating an individual reaction, but is impractical at benchmark scale: even plausible reactions can require weeks of condition optimization, and success depends on budget-dependent criteria such as yield threshold, side products, and scale, so even an experimental verdict is rarely fully objective. Within this framing, a *Safe Bet* reaction is expected to work under relatively standard conditions; *Worthwhile* and *Rather Not* reactions indicate progressively higher experimental effort, likely condition or reagent changes, and increasing risk of outright failure; and *Nonsense* reactions are those a chemist would not attempt because no reasonable starting conditions can be inferred. The scale therefore captures how useful proposed routes are under different experimental budgets: *Safe Bet* only for tight budgets, optionally including *Worthwhile* and *Rather Not* for larger budgets, and never *Nonsense*.

5 Experiments

5.1 Dataset

All of our generators and scorers are based on the proprietary Pistachio (2024Q3 release) [Mayfield et al., 2017] dataset, either used as training data (RootAligned, RP, RGP), or as a source of reference reactions (RRS). Pistachio offers substantial advantages over the commonly used USPTO-50K [Schneider et al., 2016] and USPTO-FULL [Dai et al., 2019b] datasets - it features enhanced curation, resulting in higher data quality and more comprehensive coverage of chemical reaction space. Pistachio draws from the same underlying sources (including parsed USPTO patents) with a more thorough parsing pipeline and broader reaction coverage. Because Pistachio is proprietary, we additionally retrained RP and RGP on the public USPTO-Full dataset and rebuilt the RRS reference index from the same source; these are the checkpoints we release, and the resulting scorers retain comparable performance (Appendix L). For training, we preprocess the dataset through a multi-step filtering pipeline that removes duplicate reactions, reactions from unrecognized reaction classes, entries with invalid SMILES, unmapped reactions, and reactions deemed unrelated to drug-like compound synthesis (molecules with >100 atoms, "separation" reaction classes), retaining approximately 4 million reactions.

5.2 Path-level plausibility evaluation

We compared each individual scorer (RP, RGP, RRS) and the MetaScorer at filtering implausible reactions in multi-step retrosynthesis. All variants used our BART-based RootAligned generator.

Filtering thresholds were maximized under the constraint that the system finds paths for at least 90% of targets. In practice, this corresponded to precision of 0.8 on the reaction dataset in 4, balancing strict filtering with broad target coverage. These thresholds were not tuned to optimize benchmark performance. As baselines, we report RootAligned without scoring, RootAligned with a forward reaction scorer, and LocalRetro [Chen and Jung, 2021]. The forward scorer uses the same BART architecture as RootAligned but predicts products from substrates. We used Retro* [Chen et al., 2020], following [Maziarz et al., 2025], with an expansion limit of 500 and the same eMolecules starting-material database for all systems.

Additionally, we compared against three publicly accessible retrosynthesis systems: AiZynthFinder [Genheden et al., 2020b], IBM RXN [IBM, 2025], and RetroChimera [Maziarz et al., 2024]. AiZynthFinder combines a template generator with a filtering model, IBM RXN uses a forward model in search, and RetroChimera selects top-ranked reactions from an ensemble of generators rather than an explicit reaction scorer. Access and runtime details are given in Appendix B. We deliberately chose a diverse baseline set: some systems are trained on USPTO-derived data (LocalRetro, AiZynthFinder) and others on Pistachio (RetroChimera). Despite this diversity, all of them produce a substantial fraction of *Nonsense* or *Rather Not* routes (Sec. 6.1).

We evaluated the top-1 synthesis paths for 32 targets. The 32 targets were constructed by recombining fragments of approved and investigational drugs and modifying their structures to resemble modern drugs while avoiding close literature analogs. Candidates were screened for chemical stability and molecular weight (Appendix G). Expert chemists manually evaluated each reaction using the protocol in 4; each path received the four-tier confidence score of its lowest-scoring reaction, reflecting that

one implausible step can invalidate an otherwise promising synthesis. The goal is to increase *Safe bet* paths while eliminating *Nonsense* and reducing *Rather Not* paths. While 32 may appear small relative to typical machine-learning benchmarks, each system's predicted paths contain on average $\sim$5 reactions, yielding approximately 160 expert-labeled reactions per system and well over 1,500 reactions across the full path-level comparison. All models were trained on a single A100 GPU on an internal cluster, with inference and search executed on CPU; full hardware, training-time, and runtime details are given in Appendix M.

5.3 Reaction-level plausibility prediction

We also compare individual scorers (RP, RGP, RRS) and the MetaScorer on individual reactions with expert labels from Section 4.

Model performance was assessed using precision-recall (PR) and receiver operating characteristic (ROC) curves, with PR-AUC and ROC-AUC reported for each method. Reactions rated *Safe Bet* were treated as positives. *Worthwhile* reactions were excluded as they represent borderline cases where chemist confidence is uncertain, making them neither clearly positive nor negative examples for evaluation purposes. *Rather Not* and *Nonsense* reactions were labeled as negatives. We also analyzed each failure category, reporting ROC-AUC, PR-AUC, and false positive counts.

To evaluate model complementarity, we analyzed the overlap in false positive predictions across individual scorers, calculated as:

$$\text{overlap} = \frac{\left|\bigcap_{\texttt{scorer}\in\{\texttt{RGP},\texttt{RP},\texttt{RRS}\}} \text{FP}_{\texttt{scorer}}\right|}{\min_{\texttt{scorer}\in\{\texttt{RGP},\texttt{RP},\texttt{RRS}\}} |\text{FP}_{\texttt{scorer}}|}, \tag{3}$$

where FP is a set of false positives produced by a given scorer.

6 Results

6.1 Path-level plausibility evaluation

Pathway correctness is compared in Fig. 3.

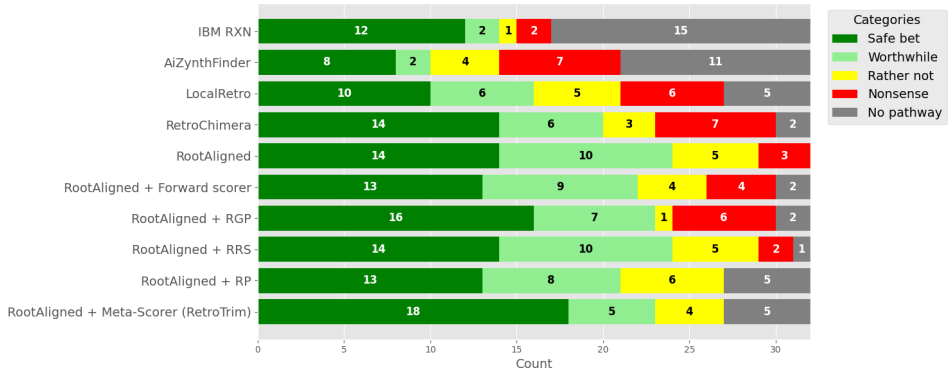


Figure 3: Comparison of RootAligned variants with IBM RXN, AiZynthFinder, LocalRetro, and RetroChimera. External baselines either miss many targets or return hallucinated routes; RootAligned without scoring has high coverage but unreliable outputs. Individual scorers improve reliability at the cost of coverage, while RetroTrim's MetaScorer gives the most trustworthy routes, maintaining zero *Nonsense* pathways while increasing *Safe Bet* results.

AiZynthFinder fails to identify viable pathways for many targets and still returns a substantial fraction of *Nonsense* or *Rather Not* routes. IBM RXN improves the number of reliable pathways and reduces hallucinations, but still fails on a considerable fraction of targets. RetroChimera has the strongest coverage among external baselines and produces many *Safe Bet* routes, yet its remaining *Nonsense*

outputs show that ensembling generators alone does not eliminate hallucinations on challenging targets.

RootAligned without scoring finds paths for all targets, but often returns unreliable *Nonsense* or *Rather Not* routes. Adding individual scorers trades coverage for reliability, a desirable trade-off when rejecting a target is preferable to returning a misleading synthesis. However, RGP, RRS, and forward scoring are insufficient on their own: they reduce unreliable routes only modestly or leave serious errors. RP is the strongest individual filter, eliminating all *Nonsense* pathways, but it also rejects more otherwise useful routes. RetroTrim's MetaScorer gives the best overall profile by combining high reliability with useful coverage: it increases *Safe Bet* paths, maintains zero *Nonsense* results, and reduces *Rather Not* pathways.

6.2 Reaction-level plausibility evaluation

Our results show that the MetaScorer outperforms individual scorers in both precision and recall, demonstrating effective integration of complementary signals. Figure 4 presents the ROC and precision-recall curves, with the MetaScorer achieving consistently higher area under the curve (AUC) values across both metrics. Similar curves broken down by reaction failure category can be found in Appendix H.

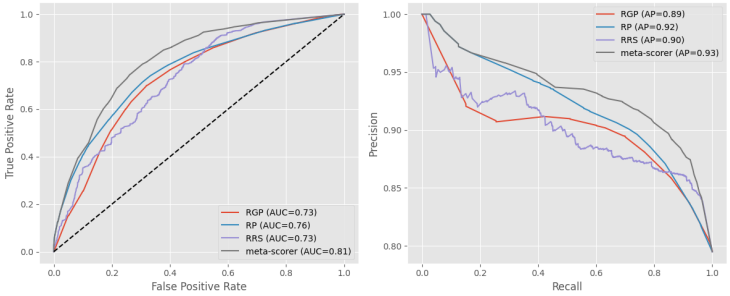


Figure 4: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer. The MetaScorer achieves higher AUC values for both ROC and PR curves, indicating better discrimination between plausible and implausible reactions. Among the individual scorers, RP shows the best performance.

Figure 5 shows ROC-AUC values for each scorer broken down into different failure categories, illustrating that individual scorers demonstrate proficiency in filtering out reactions deemed implausible under different evaluation criteria. RGP achieves the best performance on *Selectivity* and *Reactivity* errors. RRS is most capable of detecting fundamental structural issues such as *Reactant mismatches* and *Magic*, in addition to *One pot* errors. RP shows a balanced profile, which explains its overall superior performance compared to RGP and RRS in Figure 4. By leveraging the unique strengths of each individual scorer, the MetaScorer maintains robust predictive performance across all failure categories.

We also analyze the overlap between false positives that each scorer fails to filter, as shown in Figure 6. The results show distinct complementarity: while RRS and RP exhibit high overlap in most categories, it is notably reduced for *One-pot*, *Magic*, and *Reactant mismatch* failure modes — the categories where Figure 5 demonstrates RRS's superior performance. RGP and RP show consistently low overlap across all failure categories, indicating that these scorers capture different aspects of reaction implausibility. Importantly, when considering all three scorers jointly, the overlap drops to very low levels across all categories, providing strong evidence that each scorer contributes unique discriminative value essential for building a robust MetaScorer.

7 Conclusions and Limitations

In this work, we introduced RetroTrim, a retrosynthesis system that avoids hallucinations in the synthesis tree by combining three complementary reaction scoring strategies. On thirty-two unpublished drug-like targets, RetroTrim was the only compared method to completely avoid hallucinated paths

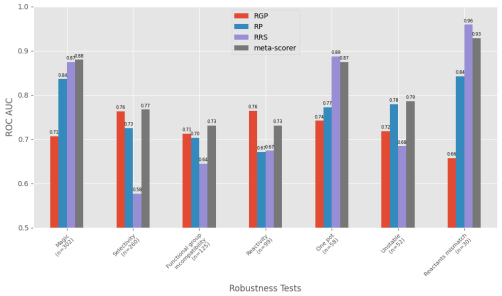


Figure 5: ROC-AUC performance of individual scorers across different failure categories, with sample sizes indicated for each category.

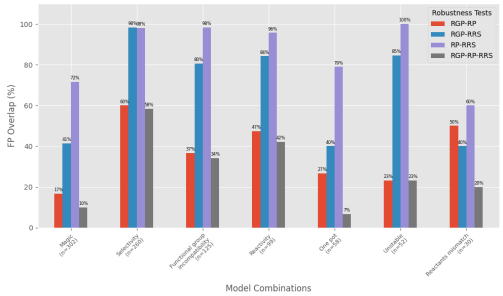


Figure 6: Overlap between individual pairs of scorers and triple of all scorers across different failure categories, with sample sizes indicated for each category.

while finding the most issue-free routes. Analysis across failure classes shows that the scorers are synergistic both for filtering individual reactions and for improving the resulting retrosynthetic paths.

In evaluating retrosynthesis systems and scorers, we made use of a novel labeling protocol where we leveraged chemists' expertise to produce fine-grained labels for generated reactions. To our knowledge, this is the first instance of such a granular analysis of retrosynthesis systems' output, where automated metrics and ad-hoc manual inspection were the norm. In order to facilitate further development in the field, we release the thirty-two targets used for path generation, the protocol, and a 500-reaction labeled subset. We also release the code of our reaction scorers, the corresponding models trained on USPTO, and their integration with the Syntheseus search framework [Maziarz et al., 2025] at `https://github.com/molecule-one/retrotrim`.

Limitations. Our evaluation relies on structured expert assessment, since experimental validation at this scale is impractical (Sec. 4). This enables fine-grained per-reaction labels backed by over 300 person-hours of PhD-level annotation. The 32 evaluated targets were deliberately selected to cover challenging, unpublished drug-like chemistry with minimal literature leakage; generalization beyond this regime remains to be shown. RetroTrim was trained on data biased toward medicinal chemistry, and we focus on top-1 paths only — the effect of plausibility filtering on path diversity is left for future work. RetroTrim was developed and evaluated for drug-like small-molecule organic chemistry; we have not validated it for cyclopeptides, polymers, inorganic compounds, or other substantially different chemical domains. Within this domain, when RetroTrim fails to return a flawless route, the bottleneck may lie in any of three places: the generator may not propose a viable disconnection, the learned scorers may be queried out of distribution, or RRS may reject rare but valid chemistry because of limited precedent. Attributing failures among these sources is left for future work. Finally, our evidence that the three scorers are complementary is empirical rather than theoretical, and they do not form a complete set of checks.

Acknowledgments and Disclosure of Funding

This work was supported in part by the European Union under the European Funds for a Modern Economy 2021–2027 programme (FENG), project no. FENG.01.01-IP.02-0907/23, "Development of machine learning algorithms and an IT system based on them to suggest chemical reaction conditions for key reaction classes".

References

Andres M Bran, Theo A Neukomm, Daniel P Armstrong, Zlatko Jončev, and Philippe Schwaller. Chemical reasoning in llms unlocks strategy-aware synthesis planning and reaction mechanism elucidation. *arXiv preprint arXiv:2503.08537*, 2025.

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*, ICML'20. JMLR.org, 2020.

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

Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science*, 3(12):1237–1245, 2017. doi: 10.1021/acscentsci.7b00355. URL https://doi.org/10.1021/acscentsci.7b00355. PMID: 29296663.

Connor W. Coley, William H. Green, and Klavs F. Jensen. Machine learning in computer-aided synthesis planning. *Accounts of chemical research*, 51 5:1281–1289, 2018. URL https://api.semanticscholar.org/CorpusID:13748494.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

Piotr Gaiński, Michał Koziarski, Krzysztof Maziarz, Marwin Segler, Jacek Tabor, and Marek Śmieja. Retrogfn: Diverse and feasible retrosynthesis using gflownets. *arXiv preprint arXiv:2406.18739*, 2024.

Samuel Genheden, Ola Engkvist, and Esben Jannik Bjerrum. A quick policy to filter reactions based on feasibility in ai-guided retrosynthetic planning. 2020a.

Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of cheminformatics*, 12(1):70, 2020b.

L.K. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

Alan Kai Hassen, Andrius Bernatavicius, Antonius P. A. Janssen, Mike Preuss, Gerard J. P. van Westen, and Djork-Arné Clevert. Atom-anchored llms speak chemistry: a retrosynthesis demonstration. *arXiv preprint arXiv:2510.16590*, 2025.

IBM. Rxn for chemistry. https://rxn.app.accelerate.science/, 2025. Accessed: 2025-08-26.

Pavel Karpov, Guillaume Godin, and Igor V. Tetko. A transformer model for retrosynthesis. In Igor V. Tetko, Věra Kůrková, Pavel Karpov, and Fabian Theis, editors, *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*, pages 817–830, Cham, 2019. Springer International Publishing. ISBN 978-3-030-30493-5.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019a. URL http://arxiv.org/abs/1910.13461.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019b. URL http://arxiv.org/abs/1910.13461.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, Thomas Hubert, Peter Choy, Cyprien de Masson d'Autume, Igor Babuschkin, Xinyun Chen, Po-Sen Huang, Johannes Welbl, Sven Gowal, Alexey Cherepanov, James Molloy, Daniel J. Mankowitz, Esme Sutherland Robson, Pushmeet Kohli, Nando de Freitas, Koray Kavukcuoglu, and Oriol Vinyals. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022. doi: 10.1126/science.abq1158. URL https://www.science.org/doi/abs/10.1126/science.abq1158.

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. URL https://doi.org/10.1021/acscentsci.7b00303. PMID: 29104927.

John Mayfield, Daniel Lowe, and Roger Sayle. Pistachio: Search and faceting of large reaction databases. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*, volume 254. AMER CHEMICAL SOC 1155 16TH ST, NW, WASHINGTON, DC 20036 USA, 2017.

Krzysztof Maziarz, Guoqing Liu, Hubert Misztela, Aleksei Kornev, Piotr Gaiński, Holger Hoefling, Mike Fortunato, Rishi Gupta, and Marwin Segler. Chimera: Accurate retrosynthesis prediction by ensembling models with diverse inductive biases. *arXiv preprint arXiv:2412.05269*, 2024.

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 Discuss.*, 256:568–586, 2025. doi: 10.1039/D4FD00093E. URL http://dx.doi.org/10.1039/D4FD00093E.

Microsoft. Retrochimera. https://ai.azure.com/catalog/models/RetroChimera, 2025. Accessed: 2025-10-10.

Anjie Qiao, Zhen Wang, Jiahua Rao, Yuedong Yang, and Zhewei Wei. Advancing retrosynthesis with retrieval-augmented graph generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):20004–20013, Apr. 2025. doi: 10.1609/aaai.v39i19.34203. URL https://ojs.aaai.org/index.php/AAAI/article/view/34203.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, Los Alamitos, CA, USA, Jun 2022. IEEE Computer Society. doi: 10.1109/CVPR52688.2022.01042. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042.

Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Dąbrowski-Tumański, Mikołaj Chromiński, Rafał Loska, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021. doi: 10.1021/acs.jcim.1c00537. URL https://doi.org/10.1021/acs.jcim.1c00537. PMID: 34251814.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. A comprehensive survey of hallucination in large language, image, video and audio foundation models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.685. URL https://aclanthology.org/2024.findings-emnlp.685/.

Nadine Schneider, Nikolaus Stiefl, and Gregory Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56, 11 2016. doi: 10.1021/acs.jcim.6b00564.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. PMID: 31572784.

Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H. Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.*, 11:3316–3325, 2020. doi: 10.1039/C9SC05704H. URL http://dx.doi.org/10.1039/C9SC05704H.

Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021. doi: 10.1126/sciadv.abe4166.

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: https://doi.org/10.1002/chem.201605499. URL https://chemistry-europe.onlinelibrary.wiley.com/doi/abs/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(7698):604–610, Mar 2018. ISSN 1476-4687. doi: 10.1038/nature25978. URL https://doi.org/10.1038/nature25978.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9405–9415. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4e2a6330465c8ffcaa696a5a16639176-Paper.pdf.

Paula Torren-Peraire, Alan Kai Hassen, Samuel Genheden, Jonas Verhoeven, Djork-Arné Clevert, Mike Preuss, and Igor V. Tetko. Models matter: the impact of single-step retrosynthesis on synthesis planning. *Digital Discovery*, 3:558–572, 2024. doi: 10.1039/D3DD00252G. URL http://dx.doi.org/10.1039/D3DD00252G.

Austin Tripp, Krzysztof Maziarz, Sarah Lewis, Marwin Segler, and José Miguel Hernández-Lobato. Retro-fallback: retrosynthetic planning in an uncertain world. *arXiv preprint arXiv:2310.09270*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, JINYU YANG, Yang Yu, and Junzhou Huang. Retroxpert: Decompose retrosynthesis prediction like a chemist. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11248–11258. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/819f46e52c25763a55cc642422644317-Paper.pdf.

Kaiyu Yang, Aidan Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan J Prenger, and Animashree Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 21573–21612. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/4441469427094f8873d0fecb0c4e1cee-Paper-Datasets_and_Benchmarks.pdf.

Yemin Yu, Ying Wei, Kun Kuang, Zhengxing Huang, Huaxiu Yao, and Fei Wu. Grasp: Navigating retrosynthetic planning with goal-driven policy. *Advances in Neural Information Processing Systems*, 35:10257–10268, 2022.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

A Examples of reaction pathways

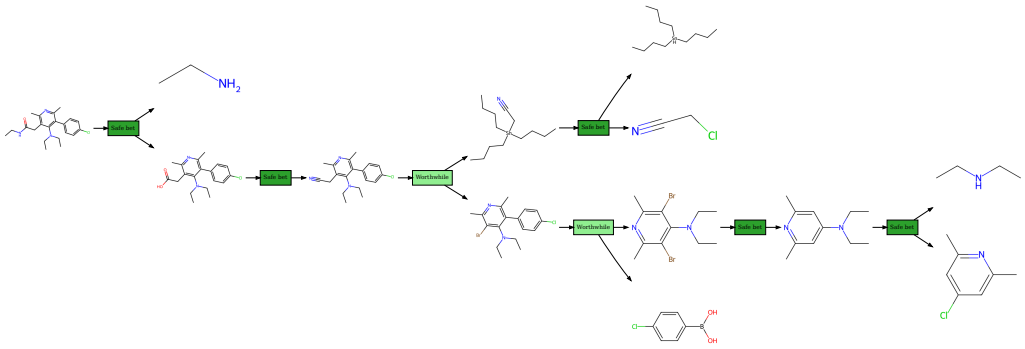


Figure 7: Example of a pathway with Safe Bet and Worthwhile reactions

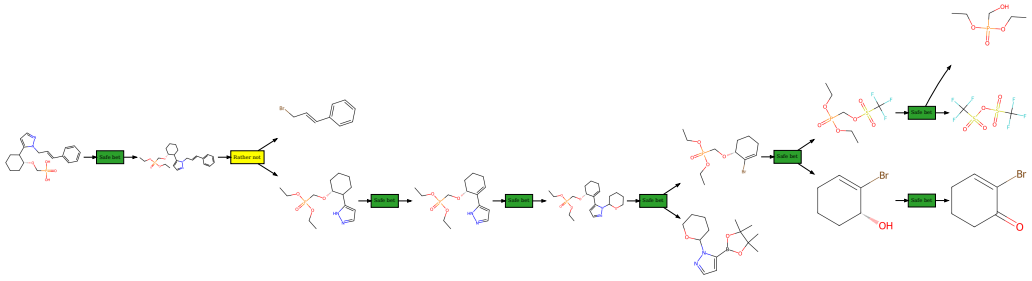


Figure 8: Example of a pathway with a Rather Not reaction

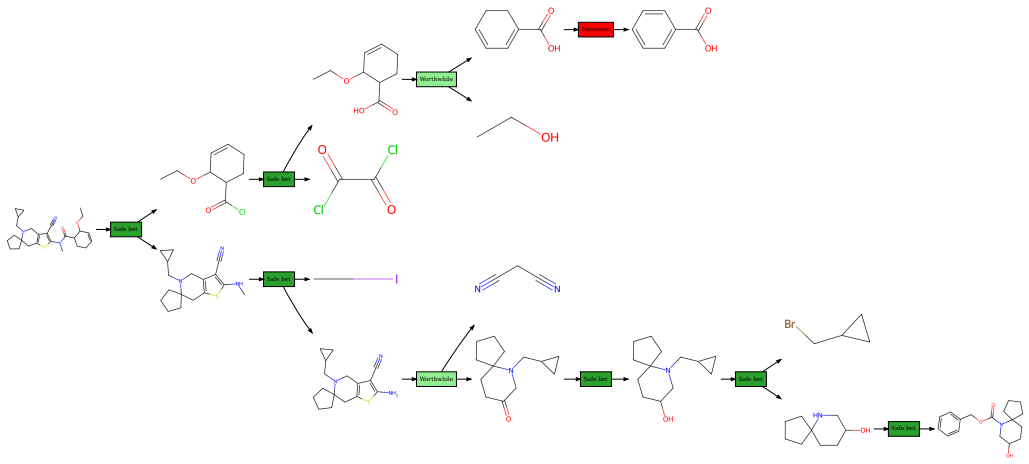


Figure 9: Example of a pathway with a Nonsense reaction

B Baseline implementation details

AiZynthFinder [Genheden et al., 2020b] was used in its default configuration from the official repository, which includes a template generator with a filtering model trained to distinguish valid reactions from artificially generated negatives; the only modification was increasing the time limit to 15 minutes to better match the runtime of other systems. IBM RXN [IBM, 2025] was accessed through its free web application. RetroChimera was queried via the Azure Foundry multi-step retrosynthesis endpoint [Microsoft, 2025]; because of a low per-call timeout, we repeated queries multiple times for each target, and prediction caching yielded an effective 15-minute search time.

C Reaction Prior details

C.1 Selection of the component weights

The architecture and training hyperparameters of the Reaction Prior model (Appendix K) were selected using the model's validation objective on held-out Pistachio splits. The component weights α, β and γ of Sec. 3.1 were then tuned on a held-out set of chemist-annotated reactions generated by our single-step models (Sec. 4), disjoint from the reactions on which the scorers are finally evaluated, using as the criterion how well the resulting RP score separates *Safe Bet* reactions from *Rather Not* and *Nonsense* ones. We were not able to improve on the default values and therefore kept them. This tuning preceded the evaluation on the 32 targets, which was run once with the configuration fixed.

C.2 Regioselectivity score

Alternative reaction sites are enumerated by applying the template extracted around the reaction center, taking the first radius in the order $2, 1, 0$ that yields productive matches. Each alternative site a is scored by the same model as the reaction itself, and compared to the desired site through the ratio of log probabilities

$$r_a = \frac{\log P_a}{\log P_{\text{desired}}}. \tag{4}$$

Alternatives with $r_a \geq 3$ are much less likely than the desired site and are discarded as non-competitive. The remaining ones contribute to a soft count

$$N = \sum_{a\,:\,r_a<3} \min(2,\ 1/r_a)\,, \tag{5}$$

so that a site the model considers at least as likely as the desired one contributes the full value of 2, while progressively less likely sites contribute progressively less. The score $S_{Regio} = 2/(1+N)$ is

then bounded in $(0, 2]$, and equals 2 when no competing site is found, which removes the need for a regularizing constant in the denominator.

D GAT

The GAT model processes reaction graphs where individual atoms and bonds are featurized with chemically-meaningful characteristics, outputting a scalar plausibility score for each reaction. The attention mechanism is modified to ensure that attention weights between non-connected nodes approach zero, preserving the chemical connectivity structure. The key difference to the original GAT is the support of global information exchange across the entire molecular graph, ensured by an artificial supernode that connects to all other nodes in the graph.

E Reference reaction scorer details

The reference count used by the Reference Reaction Scorer (RRS) is obtained from a deterministic lookup index over canonicalized reaction transformation patterns. Atom mapping is performed with RXNMapper [Schwaller et al., 2021] both for the indexed reactions (Pistachio, 2024Q3) and for the candidate reaction at query time. For each reaction, we construct two deterministic transformation keys; the same procedure is used to build the index and to compute keys for candidate reactions during inference.

Atom alignment and change detection. Reactant and product graphs are aligned by atom identity. For every aligned atom we compute an atom descriptor consisting of its element and charge state together with the sorted multiset of neighboring atom and bond descriptors. An atom is marked as changed if this descriptor differs between reactants and product. A bond is marked as changed if its bond order differs between the two sides, if it is present only on one side, or if one of its endpoints is present only on one side.

Coarse-grained key. The coarse transformation key is built from the connected component containing all changed atoms and atoms incident to changed or disappearing bonds. This subgraph is extracted from both sides of the reaction, canonicalized with a graph-canonicalization procedure that ignores the original atom order, and serialized as a reaction-pattern string. This key represents the minimal reaction-center transformation.

Fine-grained key. The fine transformation key starts from the same changed atoms and changed bonds but retains additional chemically informative context. It keeps aromatic ring bonds, non-aromatic double bonds, triple bonds, and bonds involving non-carbon or charged atoms when they are adjacent to the center or provide relevant conjugated/aromatic context. The expanded subgraph is then canonicalized and serialized in the same way as the coarse key. This makes the fine key more specific than the coarse one and separates transformations that share the same formal bond changes but occur in different conjugated, aromatic, or heteroatom-containing environments.

Index and query-time procedure. The reference index maps each coarse and fine key to the set of reference reactions that produce that key. At inference time, the candidate reaction is processed with the same procedure and the fine key is queried first; if at least one reference reaction is found, the size of the fine-key bucket is used as the reference count. If the fine-key bucket is empty, the coarse-key bucket size is used instead. The final count is therefore selected by fallback rather than by summing coarse and fine counts.

What a bucket contains. Each bucket is defined by a canonicalized local bond-edit pattern, so its members share a transformation pattern by construction rather than through a learned or approximate similarity measure. Inspecting the most populated buckets recovers established transformations such as amide formation, Suzuki–Miyaura coupling, and nitro-group reduction.

Coverage of the reference index. On the expert-labeled reactions of Sec. 4, fine-grained precedent was retrieved for 83.1% of the evaluated reactions, 8.6% fell back to the coarse-grained key, and 8.3% had no precedent at either level. For fine-grained matches, the median fine bucket contained

9,560 reference reactions (25%–75% quantiles 826–56,683), while the corresponding coarse bucket contained 34,199 (25%–75% quantiles 4,105–184,900). No reference reaction was retrieved for 42.8% of *Nonsense*, 13.4% of *Rather not*, and 3.3% of *Safe bet* reactions, so absence of precedent is strongly but not perfectly indicative of implausibility. For illustration, in a subset of the expert-labeled reaction dataset, 29 reactions passed both learned-score thresholds but had no retrieved precedent: 17 *Safe bets*, 5 *Rather nots*, and 7 *Nonsense*. The reference gate therefore trades novelty for precision: it removes implausible chemistry, but it also rejects plausible transformations that are merely uncommon in the reference corpus.

F Reaction evaluation protocol

Annotation spanned several years, with approximately five chemists involved during the main annotation periods. All annotators were salaried employees who performed labeling as part of their regular duties.

Each candidate reaction is assessed sequentially along the following dimensions. Unless specified otherwise, in each of them the reaction is scored on a four-level confidence scale: *Nonsense*, *Rather Not*, *Worthwhile*, and *Safe Bet*, indicating the plausibility of the reaction. Every reaction that is not a safe bet is assigned an additional label explaining the reason for its incorrectness.

1. **Reactant-Product Consistency:** Structural alignment between reactants and product is verified. Reactions in which the product contains a large fragment that is neither present in the substrate nor originates from a commonly used reagent, or in which no clear relationship between the atoms of the product and the substrates can be established, are marked as *Nonsense*, with the reason for incorrectness labeled as *Reactants mismatch*.

2. **Stability:** Reactions producing products or including substrates that are not isolable under the typically achievable conditions are marked as *Nonsense*, with the reason for implausibility labeled as *Unstable*.

3. **Mechanistic Plausibility:** Reactions lacking a plausible mechanism are classified as *Nonsense* or *Rather Not* due to *Magic*, covering transformations requiring unknown or highly implausible reactivity. Transformations that would require more than two non-trivial steps are also placed in this category.

4. **Multistep Feasibility and One-Pot Potential:** Reactions not achievable in a single step are assessed for decomposability into two coherent steps. If they pass this test, feasibility in a one-pot setting is scored on a four-level scale and failing reactions are marked as *One pot*.

5. **Reactivity of Substrates:** Feasibility of the reaction, given the reactivity of the substrates (e.g., electron deficiency), is verified. Reactions that cannot be reasonably expected to occur are marked as implausible, with the reason for incorrectness labeled as *Reactivity*.

6. **Functional Group Compatibility:** Molecules are screened for other functional groups that can undergo a reaction. If other groups are more probable to react first, the reaction is marked with problem *Functional group incompatibility*.

7. **Selectivity:** Selectivity of the reaction is verified, including competition between functional groups of the same type, regioisomeric outcomes (e.g., in electrophilic aromatic substitution), or other cases where multiple plausible products can arise. Reactions that fail this evaluation are marked as *Selectivity*.

F.1 Types of errors in pathways generated by evaluated systems

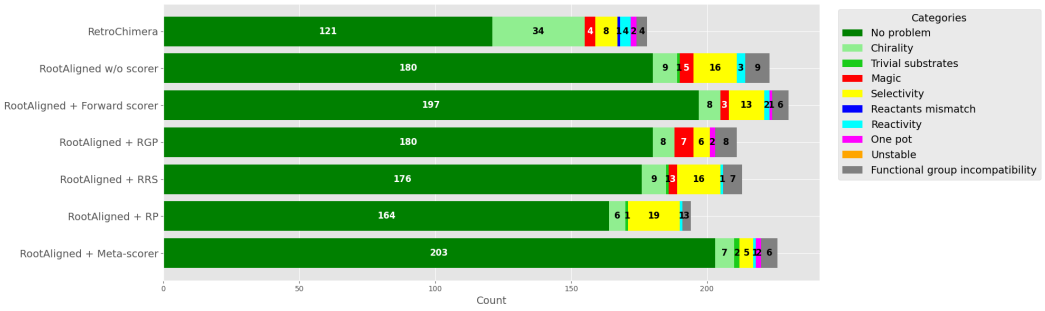


Figure 10: Detailed information on the types of errors for each reaction in the pathways from Fig. 3. Note that only reactions from the found pathways are presented here.

F.2 Implausibility annotation examples

F.2.1 Reactants mismatch

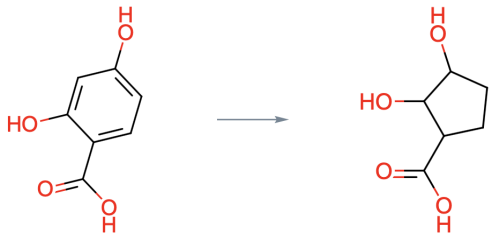


Figure 11: Nonsense: No clear relationship between atoms in the product and the substrate can be confidently proposed

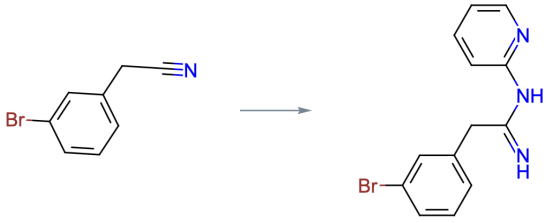


Figure 12: Nonsense: The pyridyl fragment requires an additional substrate that is missing

F.2.2 Unstable

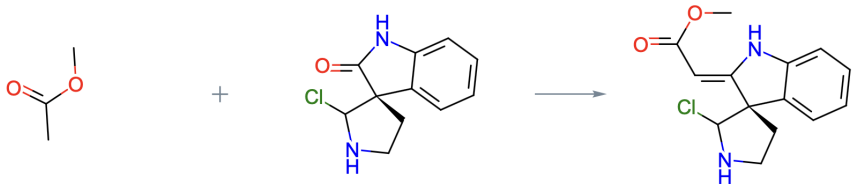


Figure 13: Nonsense: The carbon atom with amine and chlorine is not something seen in literature

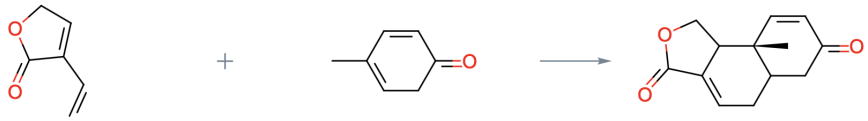


Figure 14: Nonsense: The second substrate would tautomerize to phenol instantly

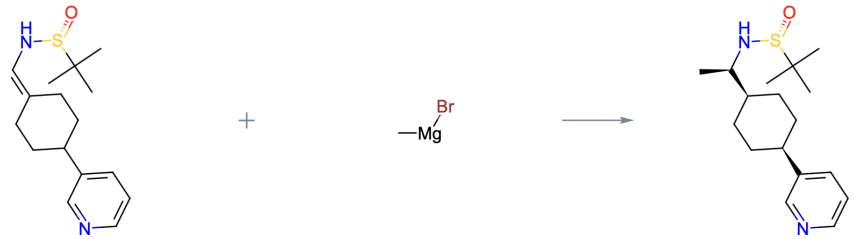


Figure 15: Nonsense: The substrate is unstable, it would tautomerize to imine

F.2.3 Magic

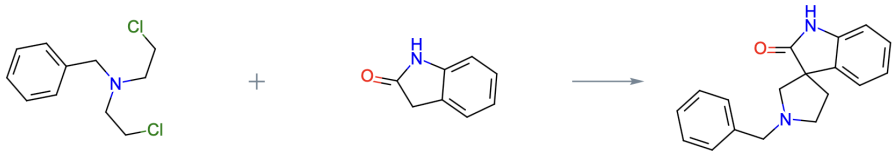


Figure 16: Nonsense: Changing length of the alkyl chain, no known precedent of such variant of carbon alkylation

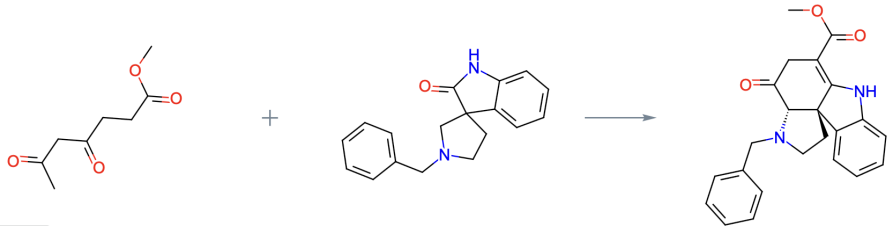


Figure 17: Nonsense: An alkyl chain acting as a leaving group and bond formation by an unactivated amine carbon. No such reactivity ever demonstrated in literature

F.2.4 One pot

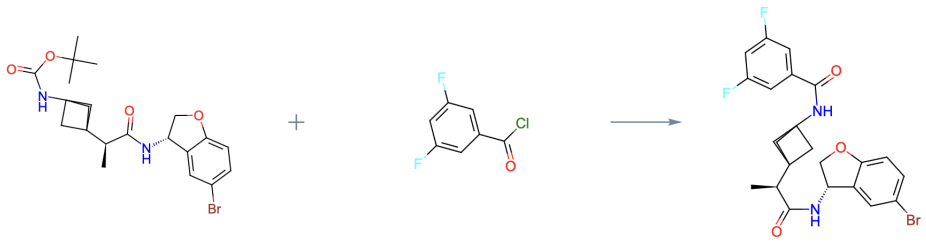


Figure 18: Rather not: 2 steps required – Boc deprotection and acylation

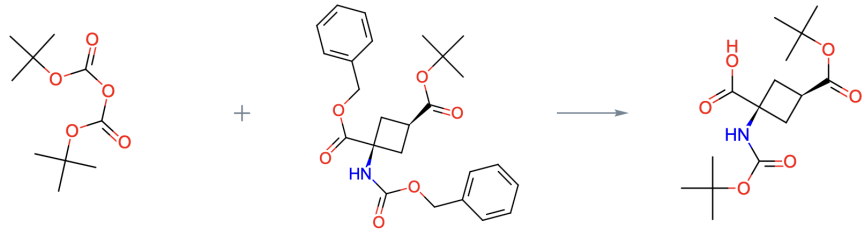


Figure 19: Rather not: 2 steps required - Cbz deprotection and Boc protection

F.2.5 Reactivity

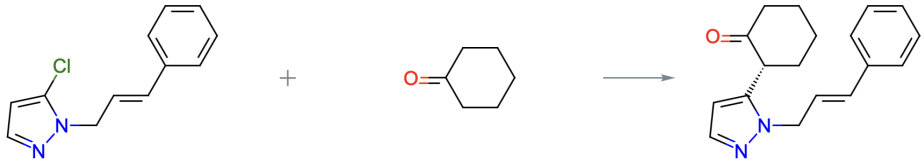


Figure 20: Rather not: Most of the references for this reaction are around electron-deficient heterocycles, only one example with pyrazole in literature

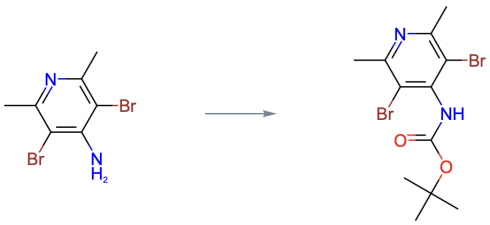


Figure 21: Rather not: High likelihood of steric hindrance

F.2.6 Functional group incompatibility

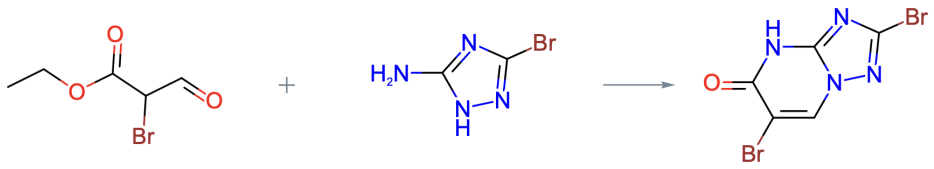


Figure 22: Rather not: No literature references where a bromine is located in alpha to the ester position. The alkyl bromine would most likely react more readily than the ester.

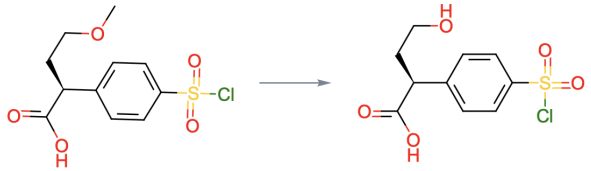


Figure 23: Nonsense: No conditions allow to cleave a methyl ether in a way that wouldn't affect the sulfonyl chloride

F.2.7 Selectivity

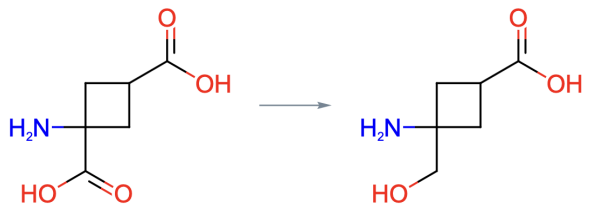


Figure 24: Rather not: There is a considerable risk that achieving the disubstituted product at a satisfactory yield would be very difficult (especially accounting for the presence of amine in the structure).

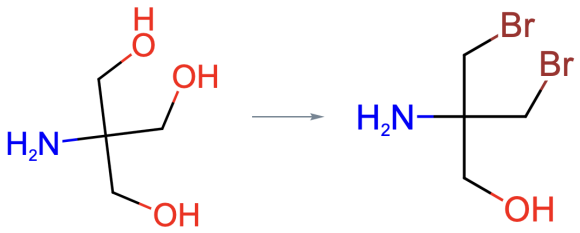


Figure 25: Rather not: There are 3 equivalent hydroxyl groups, so in bromination we expect triple substitution rather than this scenario

G Retrosynthesis targets

A PhD-level chemist manually created the targets by fragmenting approved and investigational drugs, recombining selected fragments through non-retrosynthetic connections, and making changes such as heteroatom substitution and ring expansion. The procedure aimed to produce compounds that match common features of modern drugs while avoiding close literature analogs. Candidates were screened for chemical stability and molecular weight.

G.1 SMILES

Clc1ccc(-c2c(N(CC)CC)c(c(nc2C)C)CC(=O)NCC)cc1

O(c1cc(c([N+](=O)[O-])cc1)COC1CN(C(=O)[C@@H]2C[C@]3(NC(OC3)=O)C2)C1)C1CCCC1

FC1(F)C(N2N=CC(=C2C)c2cc(ccc2)C#Cc2c(OC)cc(nc2)C(=O)O)C1

O(C(C)(C)C)[C@H](C(=O)Nc1nc2[C@](O)(CCc2cc1)CC)c1c(nc(cc1)C)C

FC1(F)Oc2c(O1)cc(nc2)C(=O)NC1=NN2C(C(=O)N[C@@H]3[C@H]2CCC3)=C1

Clc1c(N2CCC(F)(F)CC2)c(Cl)cc(NC(=O)CC[C@]2(NC(=O)NC2=O)C2CC2)c1

S(=O)(=O)(Nc1nc2N(N(C(=O)c2cn1)CC=C)C)c1ccc([C@@H](C2=Nc3c(N2)cccc3)CCO)cc1

Fc1cc(F)cc(N2[C@H](CN(CC(=O)Nc3ncnc4N(C(C)C)C=C(F)c34)CC2)C)c1

Fc1c(nc2c(c(F)ccc2)c1)Nc1cc2C(OC(=O)c2cc1)(C)C

O1C(=NN=C1)c1c(ncnc1)NC1C[C@H](O)[C@@H](O)C1

Fc1cc2c(OB(O)[C@@H](NC(=O)C3CC3)C2)cc1

S(C=1NN=NC1C(=O)NCCOCCNC(=O)C=1N=C(SC1)N1N=CC(=C1)C)c1ccccc1

O1C(Oc2c1c(ccc2C)C)([C@@H]1CC[C@@H](NC(=O)c2ncc(cc2)C#N)CC1)C

S(C1=C(C(=O)NC(=C1)C)CN(c1c2c(nccc2)c(cc1)C#N)C)C

O(CC(=O)NC1CC2N(C(C1)CC2)C)CCN1c2c3N(C(=O)C1=O)CCCc3ccc2

FC(F)(F)c1cc(C2=CN(C(=O)C(NC(=O)C3=NN(c4c3cccc4)C)=C2)C)ccc1

Fc1cc(F)cc(C(=O)NC23CC([C@@H](C(=O)N[C@H]4c5c(OC4)ccc(-c4c(OC)ccc(c4)C)c5)C)(C2)C3)c1

O1c2c(cc(C3=CN4N=C(N=C4N=C3)c3cnc(C(=O)C)cc3)cc2)CCC1

S1C(N(C(=O)C2C(OCC)C=CCC2)C)=C(C2=C1CC1(N(C2)CC2CC2)CCCC1)C#N

S(=O)(=O)(N[C@@H]([C@@H]1CC[C@H](c2cnccc2)CC1)C)c1cc(F)cc(-c2ncccc2)c1

FC(F)(F)[C@@H](N1CCC2(C(=O)N(Cc3c4OC=C(c4cc(OC(C)C)c3)C)CC2)CC1)CC1[C@@H](O)[C@@H](O)CC1

FC(F)(F)[C@@H]([C@H](C(=O)N[C@@H]([C@@](O)(N)CC)C)c1cc(OC)cc(OC)c1)C

FC(F)(F)c1ncc(-c2ncc(C(F)(F)F)c(c2)CNC(c2cc(C3=NOC(=C3CO)CC)ccc2)C2CC2)cn1

O1c2c(nc(N3C(=CC=C3C)C)nc2CCC1)NC1CCC(CO)CC1

O(c1ccc([N+](=O)[O-])cc1)CC[C@@](N)(CCN(C(=O)c1c2c(C(=O)c3c(C2=O)cccc3)ccc1)C)C

S(=O)(c1ccccc1)CCNC(=O)CN(c1ncnc([C@@H]2C[C@@H](O)C2)c1)C

Fc1c(C=2OC(=NN2)C=2Oc3c(cc4NC(Oc4c3)=O)C2)cc(F)cc1

O=C(N1C2C(Nc3ncc(-c4cnccc4)cn3)CC1CC2)C1C(O)C(O)CC1

S1[C@]2(C(=O)N3CC4[C@@H](NC5=NN(C=N5)CC(F)(F)F)[C@H](C3)CC4)[C@H]([C@](N=C1N)(c1ccccc1)C)C2

P(=O)(O)(O)CO[C@H]1C(C=2N(N=CC2)C/C=C/c2ccccc2)CCCC1

FC(F)(F)C(Nc1cncc(C(CO)C)c1)c1c(F)cc(OC2CN(C2)CCCF)cc1

O=C(N1CC(N2C(=O)CNC(C2)C)C1)N[C@H]1C(=O)NC[C@@H]1c1ccc(N2C[C@@H](O)CC2)cc1

G.2 Visualization

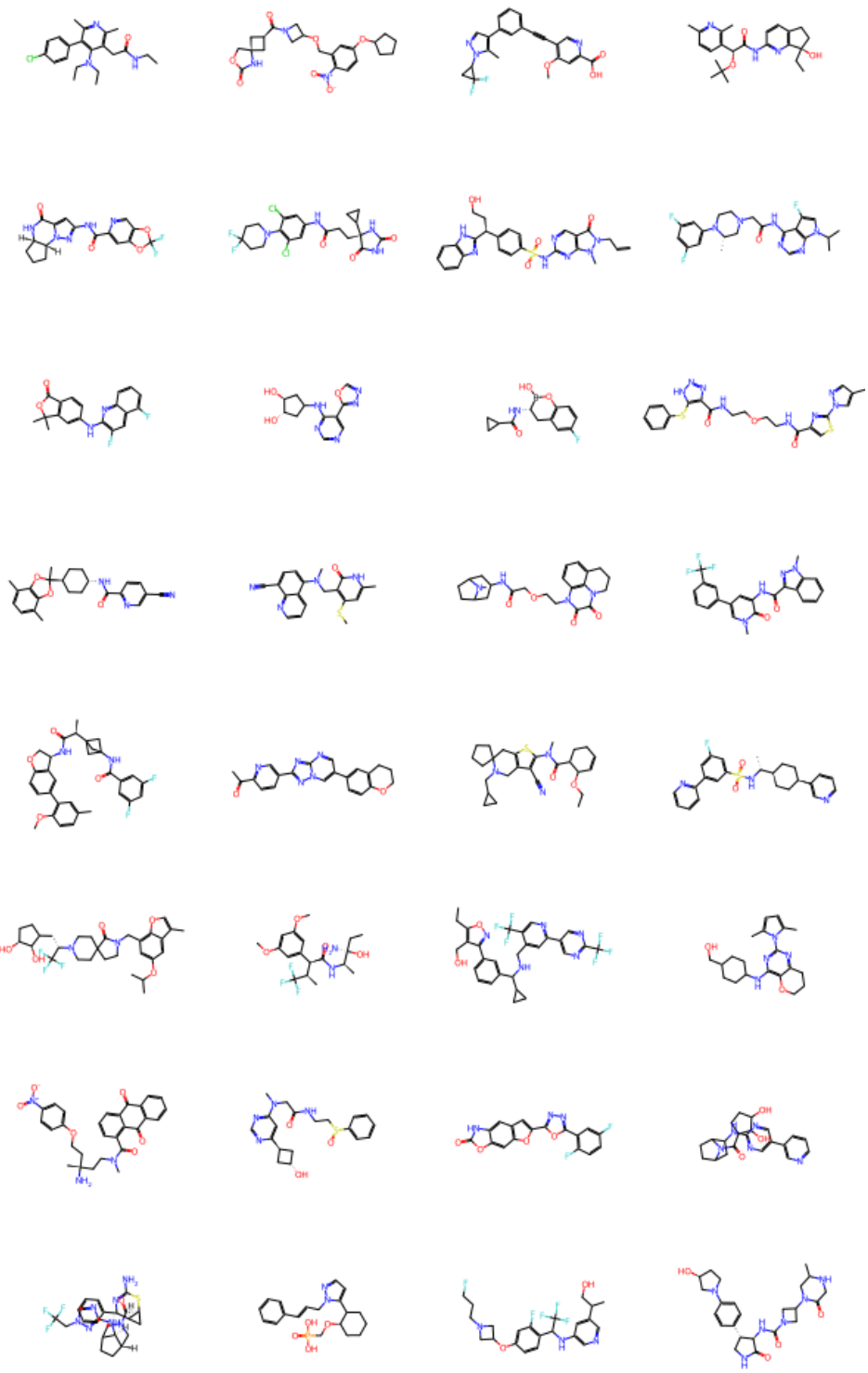


Figure 26: 32 molecules that have been used as targets for retrosynthesis.

H ROC and precision-recall curves by failure category

H.1 Magic

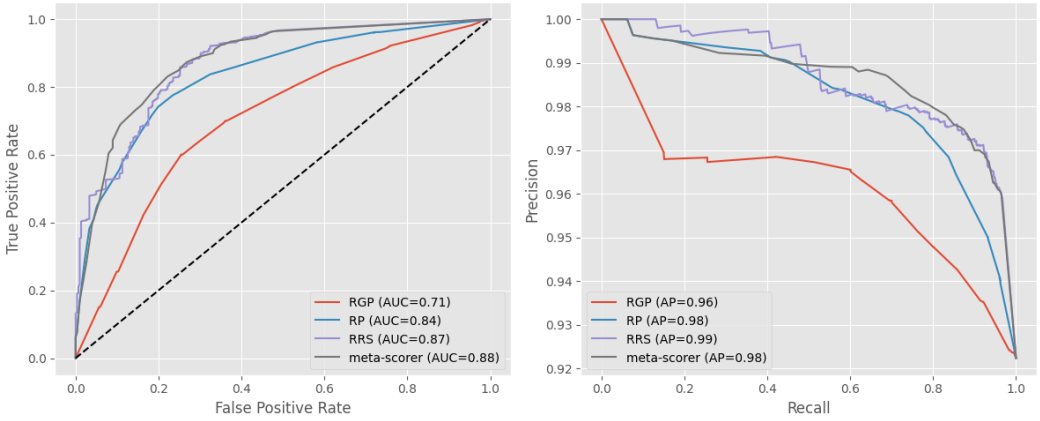


Figure 27: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Magic* and *No Problem* reactions.

H.2 Selectivity

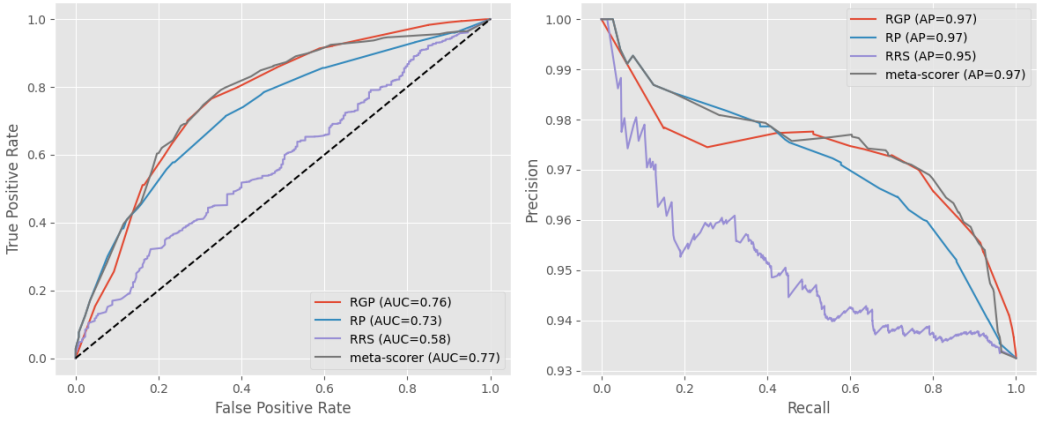


Figure 28: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Selectivity* and *No Problem* reactions.

H.3 Functional group incompatibility

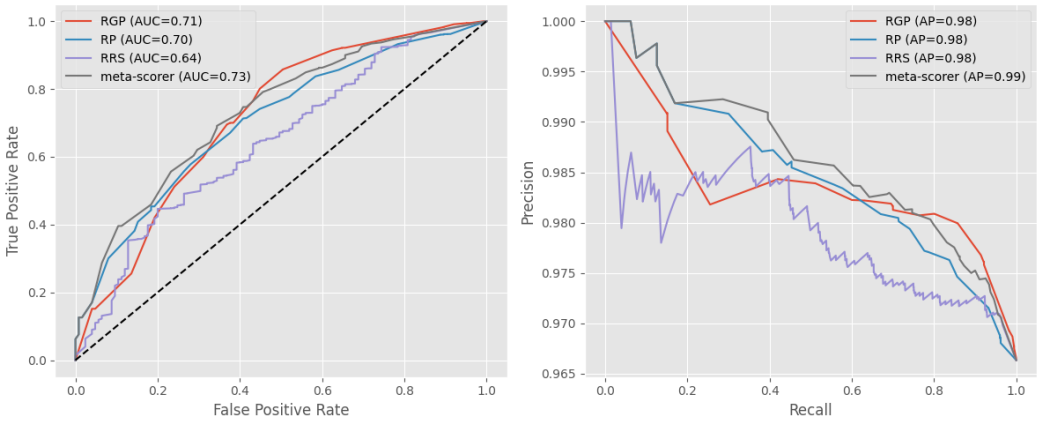


Figure 29: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Functional group incompatibility* and *No Problem* reactions.

H.4 Reactivity

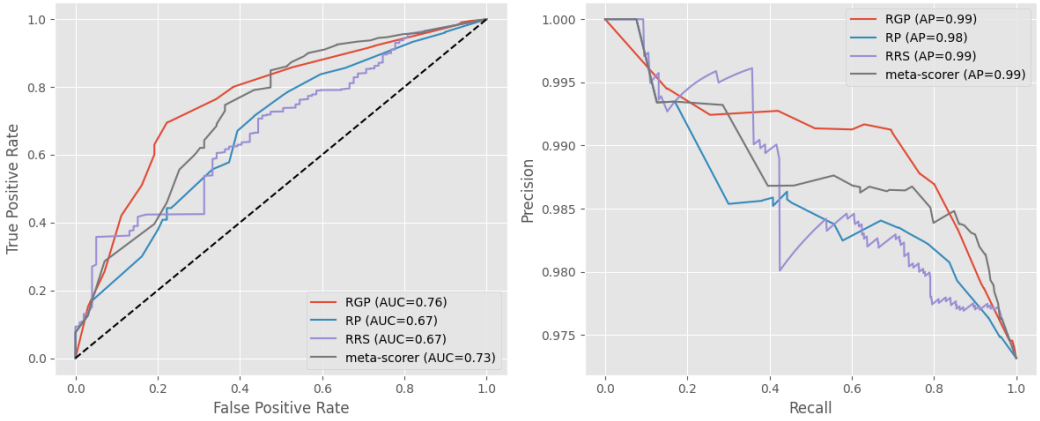


Figure 30: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Reactivity* and *No Problem* reactions.

H.5 One pot

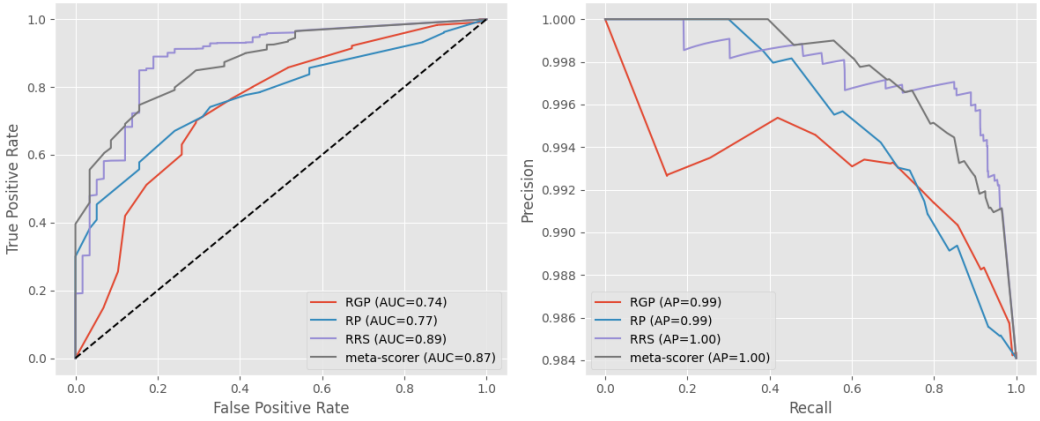


Figure 31: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *One pot* and *No Problem* reactions.

H.6 Unstable

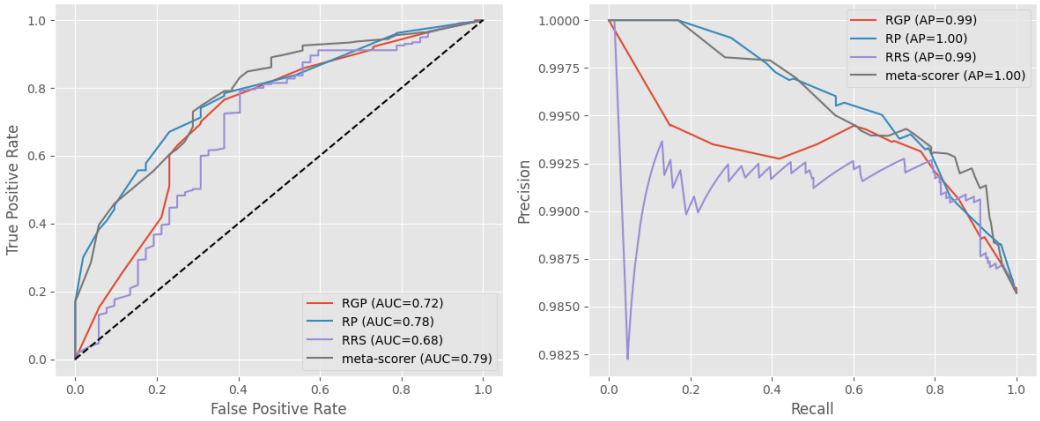


Figure 32: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Unstable* and *No Problem* reactions.

H.7 Reactants mismatch

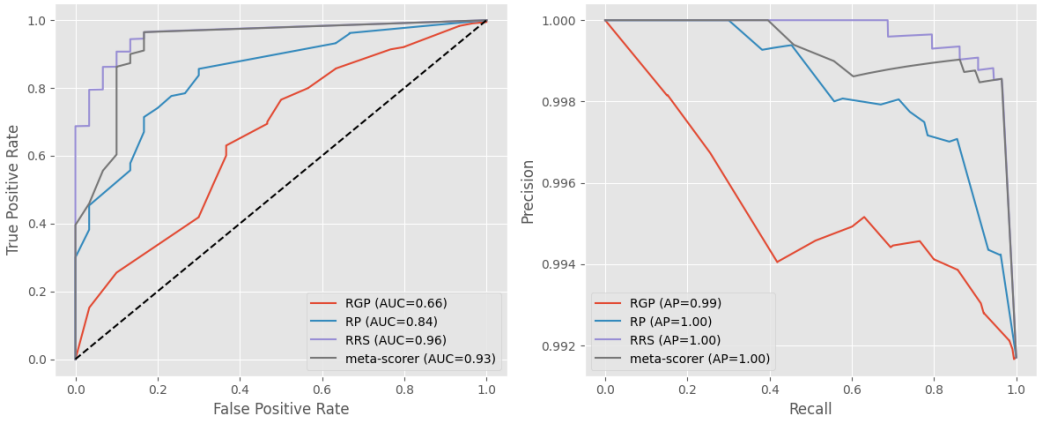


Figure 33: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the MetaScorer on *Reactants mismatch* and *No Problem* reactions.

I False positive counts by failure category

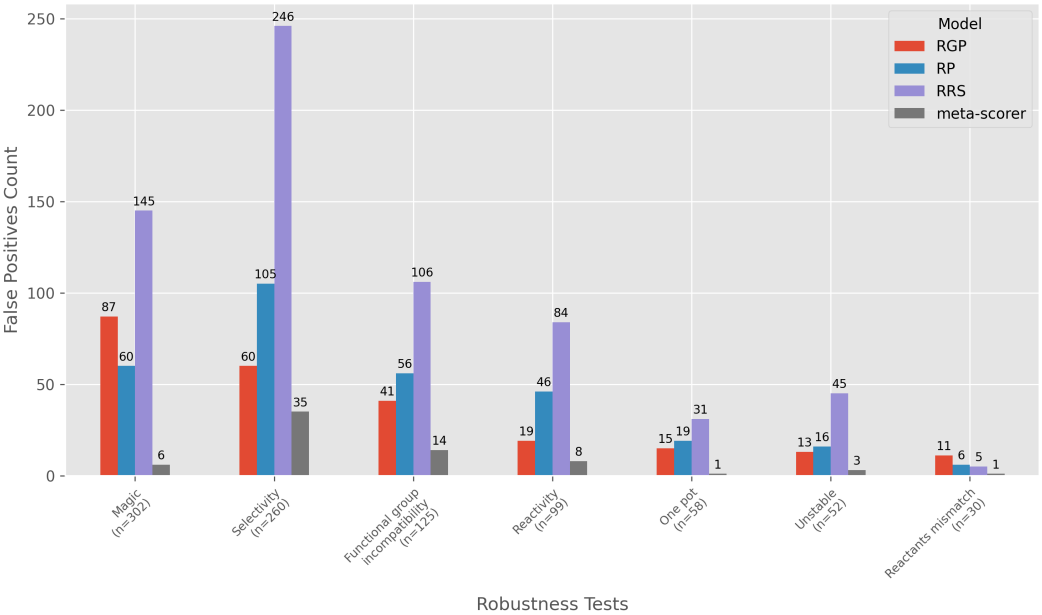


Figure 34: Counts of false positives produced by individual scorers versus the MetaScorer across different failure categories, with sample sizes indicated for each category.

J True negative counts by failure category

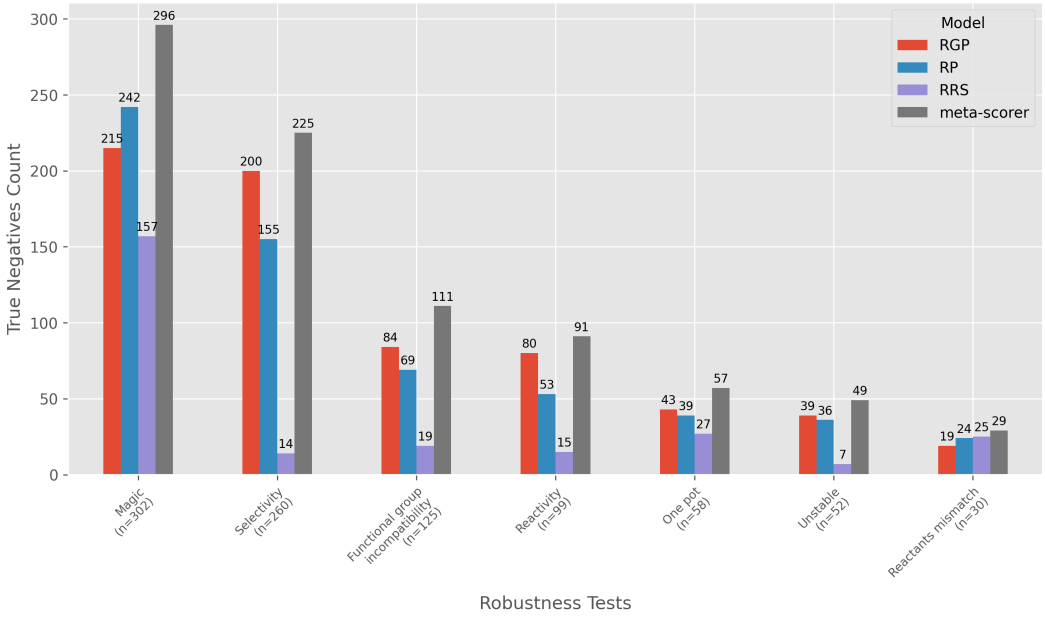


Figure 35: Counts of true negatives produced by individual scorers versus the MetaScorer across different failure categories, with sample sizes indicated for each category.

K Training hyperparameters

Table 1 summarizes the main training hyperparameters for the learned components of RetroTrim. The Reference Reaction Scorer is not trained; it uses the transformation-key lookup index described in Sec. E. MetaScorer thresholds are selected by grid search as described in Sec. 5.2.

Table 1: Main hyperparameters used for training the learned components.

Component	Hyperparameters
RootAligned generator	BART encoder-decoder with vocabulary size 256, max sequence length 512, $d_{\text{model}} = 512$, 8 encoder layers, 4 decoder layers, 8 attention heads, encoder FFN dimension 1024, decoder FFN dimension 2048, ReLU activations, dropout 0.02, attention dropout 0.05, activation dropout 0.05. Training used AdamW with learning rate $2 \cdot 10^{-5}$, weight decay 0.005, batch size 32, gradient clipping 10.0, 1000 warmup steps, cosine decay to 0.1 of the initial learning rate, and a maximum of $5 \cdot 10^6$ steps. Validation and test fractions were 5% each.
Reaction Prior	Decoder-side autoregressive scorer based on BART with vocabulary size 512, max sequence length 1024, $d_{\text{model}} = 512$, 1 encoder layer, 4 decoder layers, 1 encoder attention head, 4 decoder attention heads, encoder FFN dimension 128, decoder FFN dimension 2048, ReLU activations, dropout 0.05, attention dropout 0.05, activation dropout 0.1. Training used AdamW with learning rate $5 \cdot 10^{-5}$, weight decay 0.05, batch size 16, 16-bit precision, 1000 warmup steps, cosine decay to 0.1 of the initial learning rate, and 200,000 total scheduler steps.
Reaction Graph Plausibility	GAT with hidden dimension 1024, 8 graph-convolution layers, 3 fully connected layers, FC hidden dimension 512, residual connections, batch normalization, 8 attention heads of dimension 64, node normalization, bond-simple graph features, and a supernode. Training used BCE-with-logits loss, Adam with learning rate 10^{-3}, batch size 64, learning-rate warmup for 10,000 batches, ReduceLROnPlateau with factor 0.5 and patience 4, early stopping patience 64, unweighted positive and forward-negative examples, and generated-negative examples with weight 0.5.

Decoding settings. During search, RootAligned decodes with deterministic diverse beam search and no stochastic sampling, using `num_beams` = `num_return_sequences` = 30, `num_beam_groups` = 2, a diversity penalty of 0.025, and a maximum output length of 512 tokens. The number of groups and the diversity penalty were selected on the validation set, where diverse beam search improved top-k coverage at equal beam width. Each output is parsed with RDKit; empty or invalid outputs are discarded, and valid reactions are canonicalized and deduplicated. Discarded beams are not replaced, so each expansion returns at most 30 distinct candidates.

L Public-data reproduction on USPTO-Full

The experiments reported in the main text use Pistachio, which is proprietary. To check that the approach does not depend on it, and to make the released system usable without a Pistachio license, we retrained RP and RGP on the public USPTO-Full dataset and rebuilt the RRS reference index from the same source. The RP architecture is unchanged, with the number of training steps reduced proportionally to the smaller dataset; RGP is trained in the same manner but with a smaller architecture; RRS uses the same coarse- and fine-grained bucketing procedure with USPTO-Full as the reference source. Because USPTO-Full lacks some of the metadata available in Pistachio, in particular reaction classes, its preprocessing differs slightly from the pipeline described in Sec. 5.1.

Table 2: Reaction-level performance of RP and RGP trained on Pistachio and on the public USPTO-Full dataset. The USPTO-Full row reports the released checkpoints on the 500-reaction benchmark; the Pistachio row is measured on the chemist-annotated reactions of Sec. 4.

Training data	RP ROC-AUC	RP PR-AUC	RGP ROC-AUC	RGP PR-AUC
Pistachio	0.76	0.92	0.73	0.89
USPTO-Full	0.73	0.91	0.71	0.90

The USPTO-trained scorers retain comparable performance, indicating that the method is not specific to Pistachio. On the same 500-reaction benchmark the released RRS reaches a ROC-AUC of 0.72 and a PR-AUC of 0.89. That subset was selected to balance class proportions and issue categories, so its absolute values are not directly comparable across evaluation sets.

The released repository supports synthesis pathway search with Syntheseus [Maziarz et al., 2025], using the public pretrained RootAligned generator [Zhong et al., 2022] and our reaction scorers.

M Compute resources

Hardware. All training and evaluation was performed on an internal cluster. Model training used a single NVIDIA A100 GPU. Inference (reaction scoring) and multi-step search were run on CPU.

Training. The Reaction Prior and Reaction Graph Plausibility models were each trained for approximately 3 days on a single A100. The RootAligned generator was trained for approximately 5 days on a single A100. The Reference Reaction Scorer is not trained: it relies on a precomputed index of reference reactions keyed as described in Sec. 3.3.

Inference. Per-reaction scoring throughput on a single CPU is approximately 850 s per 1000 reactions for RP, 117 s per 1000 reactions for RGP, and 10 s per 1000 reactions for RRS. The MetaScorer runtime is effectively the sum of these (aggregation overhead is negligible). Multi-step search with Retro* takes approximately 2 hours per target on a single CPU.

Total project compute. The numbers above describe the final reported runs. Including preliminary experiments, hyperparameter exploration, and runs that were not included in the final paper, the full project consumed roughly 10× the compute reported here.

N Licenses for existing assets

External assets used in this work, with the licenses or access terms under which they were used:

- **Pistachio** (2024Q3 release) [Mayfield et al., 2017] — proprietary reaction database, used under a commercial license obtained by the authors' organization.
- **eMolecules** — starting-material database. Building blocks were obtained from the publicly available eMolecules free download. Use is governed by the eMolecules Terms of Service.
- **AiZynthFinder** [Genheden et al., 2020b] — MIT License; used as a baseline retrosynthesis system.
- **IBM RXN** [IBM, 2025] — free online tool; queried via the public web application.
- **RetroChimera** [Maziarz et al., 2024, Microsoft, 2025] — MIT License; accessed via the Azure Foundry endpoint.
- **LocalRetro** [Chen and Jung, 2021] — CC BY-NC-SA 4.0 License; used as a baseline retrosynthesis system.
- **RootAligned** [Zhong et al., 2022] — MIT License; used as the SSR generator architecture.
- **RXNMapper** [Schwaller et al., 2021] — MIT License; used for atom mapping of reactions in the reference index and at query time for the Reference Reaction Scorer.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three main contributions, each substantiated in the body of the paper: (i) among the evaluated systems, RetroTrim is the only one that eliminates all hallucinated (*Nonsense*) reactions while retaining the largest number of fully-correct routes (Sec. 6.1, Fig. 3); (ii) the MetaScorer combines three complementary scorers whose error modes are shown to be largely disjoint (Sec. 6.2, Figs. 5 and 6); (iii) we introduce a chemist-led reaction-evaluation protocol with seven issue types and a four-level confidence scale, applied to a curated benchmark of 32 drug-like targets (Sec. 4, Appendix F, Appendix G). The scope is explicitly limited to drug-like targets, matching the claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 7, covering the use of expert assessment as ground truth (and its budget-dependent interpretation), the depth-over-scale design of the 32-target benchmark, the medicinal-chemistry training bias of RetroTrim, and the restriction to top-1 paths. The interpretation of the four-tier confidence scale is detailed in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not contain theoretical results requiring proofs. Our contributions are empirical: a retrosynthesis system, a chemist-led evaluation protocol, and benchmark results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The architectures of all components (RootAligned BART generator, RP, RGP, RRS, and the MetaScorer aggregation), the Pistachio (2024Q3) preprocessing and filtering pipeline, threshold selection, the search algorithm (Retro* with MetaScorer-based filtering), and the human evaluation protocol with confidence scale and error categories are described in the main text. Additional details on baseline access and runtime (Appendix B), GAT architecture (Appendix D), reaction evaluation protocol (Appendix F), retro target SMILES (Appendix G), training hyperparameters (Appendix K), and compute resources (Appendix M) are provided. Pistachio is a commercially licensed dataset accessible to researchers under license, and the 500-reaction human-annotated subset will be released upon publication.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The 500-reaction human-annotated subset, the code of our reaction scorers, the corresponding models trained on USPTO, and their integration with the Syntheseus search framework [Maziarz et al., 2025] are released at `https://github.com/molecule-one/retrotrim` under the PolyForm Noncommercial License 1.0.0. Training data (Pistachio) is proprietary and cannot be redistributed. To support reproducibility, the architectures, preprocessing, training hyperparameters, search algorithm, threshold selection, and compute resources are documented in the main text and appendices.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The training data, preprocessing pipeline, scorer architectures, and MetaScorer aggregation are described in the Methods and Experiments sections; the search algorithm,

expansion limit, starting-material database, and threshold selection (precision ~ 0.8, $\sim 90\%$ target coverage) are stated in the Experiments section. Additional details are provided in the appendices: baseline access and runtime in Appendix B, GAT architecture in Appendix D, training hyperparameters in Appendix K, evaluation protocol and annotation effort in Sec. 4 and Appendix F, target SMILES in Appendix G, and compute resources in Appendix M.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The main quantitative comparisons (ROC/PR curves, AUC values, pathway correctness counts) are reported as single deterministic measurements on a fixed evaluation set, without error bars or confidence intervals. Pathway-level evaluation produces one annotated route per target per system, so variability across runs is not characterized.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Hardware (single A100 for training, CPU for inference and search), training times for each model, per-reaction scoring throughput, per-target search time, and an estimate of total project compute (including preliminary experiments) are reported in Appendix M.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics. Anonymity is preserved in the submission.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts: more reliable retrosynthesis planning can accelerate medicinal chemistry and drug discovery while reducing wasted experimental effort and reagent consumption on infeasible reactions. The released benchmark and evaluation protocol also enable the community to systematically diagnose and reduce hallucinations in retrosynthesis systems. Negative impacts: like all retrosynthesis tools, the underlying capability could in principle assist in planning the synthesis of harmful compounds (e.g., controlled substances, toxins). Our work does not provide capabilities beyond those already accessible through existing public retrosynthesis systems (AiZynthFinder, IBM RXN), and we release a 500-reaction labeled benchmark and our reaction scorers, not trained generative models. A secondary risk is over-reliance on confidence scores by non-expert users; the four-tier scale is presented as a utility scale that requires expert interpretation rather than a guarantee of correctness.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released assets (a 500-reaction benchmark of labeled retrosynthetic intermediates and an evaluation protocol) pose no significant risk of misuse. No pre-trained generative models with dual-use potential are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets used in this work (Pistachio, eMolecules, AiZynthFinder, IBM RXN, RetroChimera, LocalRetro, RootAligned, RXNMapper) are cited in the relevant sections, and their licenses or access terms are listed in Appendix N.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: We release three new assets: a benchmark of 32 unpublished drug-like targets (SMILES and visualizations in Appendix G), a chemist-led reaction evaluation protocol with annotated examples (Appendix F), and a 500-reaction labeled subset of the curated reaction dataset (Sec. 4). The annotation protocol, error categories, scoring scale, and per-reaction labeling effort are documented in the paper. We further release the code of our reaction scorers, the corresponding models trained on USPTO, and their integration with the Syntheseus search framework [Maziarz et al., 2025] at `https://github.com/molecule-one/retrotrim`. The released assets, including the 500-reaction labeled dataset, are distributed under the PolyForm Noncommercial License 1.0.0.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects research. Reactions were annotated by chemists employed by the authors' organization as part of their regular duties; details of the annotation effort are given in Sec. 4.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The study does not involve human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not an important, original, or non-standard component of the core methods proposed in this paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.