Trustworthy Retrosynthesis: Mitigating Hallucinations with Reaction Plausibility Filtering and Retrieval-Augmented Scoring

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

Abstract

While Artificial Intelligence (AI) has lead to significant improvements in Computer-Aided Synthesis Planning (CASP), its credibility within the chemical community is fragile. AI retrosynthesis models frequently "hallucinate" chemically implausible reactions, which undermines trust. To address this, we propose a framework that integrates three orthogonal validation strategies to ensure reaction plausibility. This key insight, combining reaction validation strategies which cover different error patterns, was the basis of our winning solution to the *[Anonymized for the sake of blind review]* Retrosynthesis Challange. Our approach combines: (1) a novel Transformer-based model, called Reaction Prior, that estimates reaction likelihood from large-scale experimental data, mimicking chemical reasoning (2) a Graph Neural Network trained on a reaction dataset augmented with synthetically generated incorrect reactions, and (3) a retrieval-based scoring system that leverages chemical databases and grounds suggestions in known chemical literature. The framework was validated on unseen targets through a novel human evaluation process, successfully rejecting the most hallucinated reactions. In this evaluation, chemical experts manually reviewed reactions within the synthetic paths, providing a more reliable and trustworthy form of verification compared to purely automatic methods.

1 Introduction

In chemistry, synthesis planning is the process of constructing synthetic routes - sequences of chemical reactions - that lead to a desired target molecule. Usually, this is done through retrosynthetic analysis, which works backward from the target, looking for suitable precursor molecules. This single-step retrosynthesis (SSR) procedure is repeated in a recursive fashion until a satisfying synthetic route is found from the available starting materials (building blocks) to the target.

Computer-Aided Synthesis Planning (CASP) leverages computational tools and algorithms to assist chemists in designing synthetic routes for chemical compounds. Today, this area is undergoing a major transformation driven by advances in Artificial Intelligence (AI) and is becoming increasingly important in drug discovery, agrochemicals, and materials science [5, 9, 7]. Unlike traditional methods that rely heavily on expert intuition or rigid rule-based systems, AI-driven CASP aims to

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

automatically learn chemical knowledge from experimental datasets. This data-centric approach addresses key limitations of previous methods, such as dependence on scarce human expertise and the high computational costs associated with exhaustive search [15]. However, it is not without its flaws.

Modern CASP systems typically generate complete synthetic routes by iteratively applying predictions from SSR model within a graph-based search algorithm. Consequently, the quality of the final synthesis plan depends on the accuracy of the underlying SSR model [32]. Despite significant progress, a critical challenge remains: these models frequently "hallucinate" proposing chemically nonsensical or implausible reactions [26]. This leads to serious efficiency concerns as each wrong prediction initiates the exploration of an invalid branch in the synthesis tree, wasting computational resources.

Beyond computational cost, hallucinations undermine the credibility of AI-driven CASP. A synthetic route is only as strong as its weakest link: A single flawed step can compromise the entire plan. When suggested routes require extensive manual checking to filter out chemically implausible steps, the systems' credibility as a whole is undermined.

Complicating matters further, "hallucination" errors are not only frequent but also highly diverse. These errors can range from obvious flaws, such as atoms appearing without a valid chemical origin, to violations of stereochemical or regioselective rules. This variety makes it difficult to overcome with any single validation strategy. One straightforward idea is to filter pathways using a high threshold on thegenerator (SSR) score, if it is explicitly modeled. However, this approach often discards pathways entirely rather than flagging nuanced problems. Alternative methods exist, such as using forward-prediction models to check reversibility, but they are typically used in isolation. Consequently, they provide only partial mitigation and remain inadequate against the full spectrum of errors. Therefore a more robust approach is needed.

In this work, we address the challenge of identifying implausible reactions during retrosynthetic planning. We propose a new framework that combines three complementary validation strategies into a unified scoring mechanism (Meta-Scorer) which identifies and removes implausible reactions generated during SSR. We show that each component of our framework specializes in different types of errors, providing broader and more reliable coverage than any single strategy could achieve alone.

Our approach is based on the principle of ensemble learning: combining diverse models with distinct error patterns yields a more robust and accurate assessment of reaction plausibility. The resulting aggregated plausibility scores are designed to be directly integrated into heuristic-guided search algorithms, such as Retro*. This enables implausible branches to be dynamically pruned from the synthesis search tree and reaction costs to be adjusted, guiding the planner towards synthetic routes that are trusted by chemical experts. This approach contrasts with prior efforts that have typically relied on a single validation strategy. Our retrosynthesis systems, combining multiple validation strategies, won the the *[Anonymized for the sake of blind review]* Retrosynthesis Challange, in which the correctness of pathways judged by human experts was crucial.

In order to evaluate our system, we employed for the first time a structured reaction plausibility validation protocol, in which PhD-level chemistry specialists assessed each reaction in the proposed retrosynthetic pathways according to a predefined labelling schema. Reactions without significant issues were labeled *No problem*, while those with errors were assigned to one of the following categories: *Magic*, *Selectivity*, *Functional group incompatibility*, *Reactivity*, *One pot*, *Unstable*, or *Reactants mismatch*. Based on the severity of any identified errors, each reaction was then given a confidence label on a four-point scale: *Nonsense*, *Rather not*, *Worthwhile*, or *Safe bet*. Then the overall confidence for a synthetic pathway was determined by the lowest-scoring reaction within it. The complete protocol is detailed in the Appendix B.

Our key contribution is a complementary set of validation strategies that effectively eliminates most severe hallucinated reactions during retrosynthetic search - labeled *Nonsense* in our protocol - and significantly reduces *Rather not* and *Worthwhile* reactions within pathways. The framework is built around a central, best-performing scorer based on novel scoring mechanism, which we call Reaction Prior, supported by two additional scorers designed to be compliment it by targeting specific error types. The framework includes:

1. Reaction Prior (RP): The core novel model, which estimates the intrinsic plausibility of reactions using large-scale experimental data. It is based on a Transformer-based architecture and designed to mimic how chemists think when evaluating reaction plausibility.

2. Graph Attention Network (GAT): A model trained to distinguish valid reactions from synthetically inplausible negatives and specializes in *Selectivity* and *Reactivity* error types.
3. Reaction Retrieval (RR): A mechanism that assesses the similarity of proposed reactions to known experimental precedents in reaction databases. It is designed to handle *Reactants mismatch*, *One pot Magic* error types.

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 the limited work on the techniques used for their validation.

2.1 Reaction Generation: Single-Step Retrosynthesis

Template-Based Methods operate by matching a target molecule against a predefined library of reaction templates, which are abstract representations of known chemical transformations. Examples include RetroSim [4], NeuralSym [20], GLN [6], DualTB [24] LocalRetro [2], RetroKNN [29] and RetroComposer [31]. The primary advantage of these models is their ability to produce interpretable outputs for reactions that fall within their template library. Though often reliable, they are not immune to proposing implausible reactions. The construction of these template libraries is challenging and involves an trade-off between specificity and computational cost. This leads to trivial errors, over-generalizations, or missing context in the template definitions, leading to chemically incorrect reaction suggestions produced by these methods [5].

Template-Free Methods aim to learn the underlying rules of chemistry directly from reaction data. These approaches typically employ machine learning architectures, such as sequence-to-sequence models [14], Transformers [12, 11], or Graph Neural Networks (GNNs) [18]. A key benefit of these approaches is their scalability and potential to generalize to novel chemical transformations not seen in template libraries [32]. However, this flexibility often comes at the cost of chemical correctness. The models may struggle with correct atom-mapping during generation or overlook crucial 2D and 3D structural information [32], which lead to predictions that violate established chemical principles and steric constraints [33].

Semi-Template-Based Methods seek to combine the reliability of templates with the flexibility of template-free generation. This category includes methods like GraphRetro [23], RetroXpert [30], and G2G [22]. While these hybrid approaches often improves performance by balancing constraints with flexibility, they do not fully eliminate the problems of generating implausible predictions.

2.2 Reaction Validation

To address the shortcomings of SSR models, researchers have developed various post-hoc validation strategies, where proposed reactions are scored and filtered to enhance the overall trustworthiness and reliability of the predictions.

Learning-Based Plausibility Models are techniques such as "in-scope filter," a neural classifier trained to predict if a given reaction is valid [21]. The ASKCOS platform implements this idea: an internal “fast filter model” scores each candidate reaction’s plausibility and reactions below a likelihood threshold are removed. These filters are typically trained using contrastive learning on real vs. bad (mismatched) reaction pairs. One example includes HiCLR framework [28], which trains a model to distinguish between reactions sharing a common chemical superclass and those that do not. An alternative strategy is to use Molecular Transformer architecture as it has been shown that its confidence in a prediction, derived from output token probabilities, correlates well with the reaction’s correctness [19].

Evidence-Based Validation via Retrieval grounds model predictions in established chemical knowledge through retrieval-augmented methods. This mirrors a chemist’s workflow of searching for literature precedents. For example, the Retrieval-Augmented RetroBridge (RARB) framework retrieves similar molecules from a database to guide the generation of reactants [17]. The rise of

LLMs has also seen the application of Retrieval-Augmented Generation (RAG), where the model's output is conditioned on retrieved documents to improve factual accuracy [25].

2.3 Integrating Validation into Synthesis Planning

Ultimately, the goal of a plausibility score is to improve the reliability of reactions and enhance multi-step planning. In this context, scores are integrated into graph-based search algorithms like Retro* or Monte Carlo Tree Search (MCTS) to guide the construction of the synthesis tree [5]. The plausibility score of a predicted invalid reaction is often used to increase the cost of a synthesis branch depending on the severity of predicted error, thus discouraging the search algorithm from exploring pathways containing that step. An effective validation method can "prune" entire branches of the search tree that begin with an invalid reaction, saving significant computational resources. However, the effectiveness of this pruning is entirely dependent on the reliability of the underlying correctness score. Relying on a single validation strategy, especially one with known weaknesses, limits the search algorithm's ability to consistently find valid and optimal routes. This highlights the need for a more robust validation approach, which is the central contribution of our work.

3 Methods

The framework is built around a central, best-performing scorer based on novel scoring mechanism that mimics intuition of chemical experts, which we call Reaction Prior, supported by two additional scorers designed to compliment it by targeting specific error types. The final framework is aggregated into a Meta-Scorer that provides final, robust assesment of reaction correctness.

3.1 Reaction Prior

The Reaction Prior (RP) is a novel sequence model trained on SMILES-formatted reactions to estimate the joint likelihood of substrates and products. Its design principle is to align the model's predictive biases with the chemical intuition that experienced chemists use to evaluate the correctness of a reaction. Reaction Prior uses a standard autoregressive, encoder-decoder BART [13] architecture, where substrates and products are processed by the decoder and trained to maximize the probability of the correct next token using a standard cross-entropy loss function. In order to produce a reaction score, Reaction Prior integrates multiple chemical considerations: overall feasibility, regioselectivity, and reactive site confidence. The final score (S_{final}) is a weighted combination of three components: $S_{final} = S_{RP}^{\alpha} \cdot S_{Regio}^{\beta} \cdot S_{RC}^{\gamma}$. Here, S_{RP} is the Reaction Prior Score, S_{Regio} is the Regioselectivity Score, and S_{RC} is the Reaction Center Score, with α, β, and γ serving as weighting factors.

Reaction Prior Score (S_{RP}) This score reflects the overall feasibility of the reaction. It is the log probability assigned by the transformer model to the complete reaction sequence, normalized by the square root of the total number of tokens (T): $S_{RP} = \frac{1}{\sqrt{T}} \log P(\text{reaction})$.

Regioselectivity Score (S_{Regio}) This component quantifies reaction site specificity. It is calculated by comparing the probability of the desired reaction (P_{desired}) to the summed probabilities of all alternative reactions at different potential sites ($P_{\text{undesired}}$): $S_{Regio} = \log\left(\frac{P_{\text{desired}}}{P_{\text{undesired}}+\epsilon}\right)$, where ϵ is a small constant to prevent division by zero.

Reaction Center Score (S_{RC}) This score evaluates the model's confidence in the identified reactive sites. It is the sum of log probabilities for the tokens representing atoms at the reaction's core, normalized by the number of such tokens (T_{RC}): $S_{RC} = \frac{1}{T_{RC}} \sum_{i \in \text{reaction center}} \log P(\text{token}_i)$.

3.2 GAT-Based Plausibility Classifier

A Graph Attention Network (GAT) [27] is trained to differentiate chemically valid reactions from implausible ones. Training uses positive examples from curated datasets and synthetic negative examples generated through forward and two-step backward template applications.

The 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.

3.3 Reference Reaction Retrieval Scorer

Retrieval-based approaches assess reaction plausibility by comparing candidate reactions against a comprehensive database of validated chemical transformations. This comparative framework evaluates how closely predicted reactions align with established chemical precedents.

We developed a structured retrieval pipeline that extracts chemical precedent information through a two-tiered reaction clustering framework based on bond change patterns. First *Coarse-grained clustering* extracts connected components of the reaction center and applies atom mapping to identify the underlying transformation pattern. Reactions belong to the same cluster if their transformation patterns are identical. Then *Fine-grained clustering* extends the coarse-grained approach by incorporating chemically significant substructures such as aromatic systems and conjugated double bonds - into the cluster classification.

Our Reference Reaction Retrieval Scorer (RR) quantifies reaction plausibility through a logarithmic transformation of the unique reference reaction count within the candidate reaction's coarse-grained cluster and fine-grained cluster:

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

where $n_{ref}(reaction)$ represents the unique number of reference reactions in the coarse-grained and fine-grained clusters containing $reaction$.

3.4 Meta-Scorer Aggregation

To improve reaction filtering, our Meta-Scorer integrates scores from Reaction Prior (RP) and GAT models, and empirical precedents ($n_{\text{ref}} > 0$) retrieved via the pipeline in Sec. 3.3. This hybrid approach mitigates the weaknesses of purely data-driven or precedent-based methods. The continuous score is described by equation $\text{score}_{\text{META}} = \max(\text{score}_{\text{GAT}}, \text{score}_{\text{RP}})$ if $n_{\text{ref}} > 0$ (0 otherwise).

For binary classification tasks and search, reactions are filtered using predefined thresholds, selected through grid search to balance the recall and precision:

$$\text{score}_{\text{META}} = \begin{cases} 1 & \text{if score}_{\text{GAT}} > 0.85 \text{ and score}_{\text{RP}} > 0.75 \text{ and } n_{\text{ref}} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

By synthesizing diverse evidence types Meta-Scorer enables more reliable reaction filtering for multi-step synthesis planning. This integrated approach mitigates individual weaknesses of purely data-driven or precedent-based methods, yielding improved performance.

3.5 Integration with Search (Retro*)

The calibrated Meta-Scorer is used during multi-step retrosynthesis search to improve the quality of predicted pathways. We integrate it into the Retro* [1] search framework as a reaction filtering mechanism. Reactions are pruned from the search tree before any further expansion if the $\text{score}_{\text{META}}$ in 2 is equal to 0. This eliminated low-quality reactions early, reducing the search space and improving overall plausibility. This integration made our system more robust to hallucinations from the underlying SSR model and helped produce more correct, trustworthy synthesis plans.

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 one of the expert-defined categories, creating the first comprehensive dataset of its kind. This resource provided a robust way to evaluate the error patterns of our reaction plausibility scorers.

Reaction Evaluation Protocol was designed 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*. For the system to be reliable, the paths must consist primarily of reactions that the experts consider to be "safe bets". On the other hand, *Nonsense* and *Rather Not* reactions can undermine user confidence and trust. Reactions that do not pass evaluation at a certain stage receive an additional label specifying the cause of their incorrectness, out of *Reactants mismatch*, *Unstable*, *Magic*, *One pot*, *Reactivity*, *Functional group incompatibility*, and *Selectivity*. Otherwise, a reaction is assigned a *No Problem* label with *Safe Bet* confidence. A detailed description of the evaluation framework is provided in Appendix B.

5 Experiments

5.1 Reaction-Level Plausibility Prediction

We compared the ability of each individual scorer (RP, GAT and RR) and the Meta-Scorer to distinguish correct from incorrect reactions. All models were evaluated on a held-out dataset of reactions derived from retrosynthesis paths for unseen molecular targets. In order to improve the alignment of the scores with probabilities of a reaction being correct, GAT and RP were calibrated using isotonic regression on a sample of representative data. Ground truth labels were established through expert chemist evaluations descibed in Section 4.

Model performance was assessed using precision-recall (PR) and receiver operating characteristic (ROC) curves, with area under the curve metrics (PR-AUC and ROC-AUC) reported for each method. Reactions with confidence rating *Safe Bet* or *Worthwhile* were treated as positive examples, while all others were labeled as negatives. We also conducted additional analysis across specific failure categories: *Magic*, *Selectivity*, *Functional group incompatibility*, *Reactivity*, *One pot*, *Unstable*, and *Reactants mismatch* reporting false positive rates and ROC-AUC scores.

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

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

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

5.2 Path-Level Plausibility Evaluation

In addition to evaluating individual reactions, we assess the plausibility of entire retrosynthetic paths. We evaluated a set of top-1 paths generated for selected 32 targets (can be found in C) by various retrosynthesis systems, both with and without our reaction filtering mechanism. Paths are assigned the same four-tier confidence score as reactions (*Safe Bet*, *Worthwhile*, *Rather Not*, *Nonsense*), determined by the lowest-scoring reaction in the path. This conservative scoring reflects the intuition that a single implausible step can invalidate an otherwise promising synthesis. Paths marked as *Safe Bet* represent routes where experts have confidence in every reaction step - increasing their proportion, along with eliminating *Nonsense* and decreasing number of *Rather Not* constitutes a fundamental requirement for reliable retrosynthesis systems.

6 Results

6.1 Reaction-Level Plausibility Evaluation

Our results show that the Meta-Scorer outperforms individual scorers in both precision and recall, demonstrating effective integration of complementary signals. Figure 1 presents the ROC and precision-recall curves, with the Meta-Scorer 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 D.

Figure 2 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

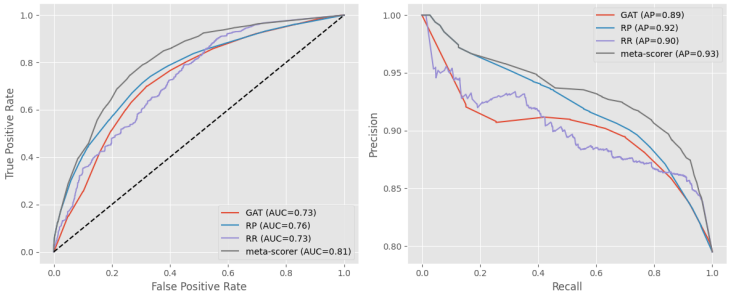

Figure 1: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer. The meta-scorer achieve higher AUC values for both ROC and PR curves, indicating better discrimination between plausible and implausible reactions. Among the individual scorers, RP shows the best performance.

under different evaluation criteria. GAT achieves the best performance on *Selectivity* and *Reactivity* errors. RR 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 GAT and RR in Figure 1. By leveraging the unique strengths of each individual scorer, the Meta-Scorer 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 3. The results show distinct complementarity: while RR 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 2 demonstrates RR's superior performance. GAT 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 Meta-Scorer.

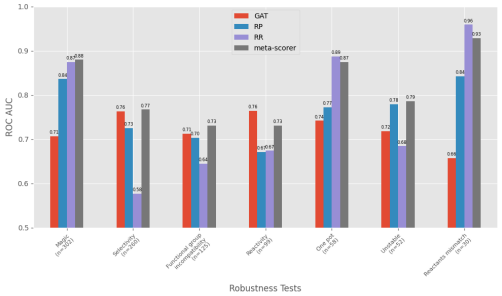

Figure 2: ROC-AUC performance of individual scorers across different failure categories, with sample sizes indicated for each category. The results reveal complementary strengths among scorers: GAT demonstrates superior performance for *Selectivity* and *Reactivity* errors, RR excels at detecting *Reactant mismatches*, *One pot* errors, and *Magic* reactions, while RP shows the highest performance for *Unstable* reactions and maintains consistent generalist performance across other failure types.

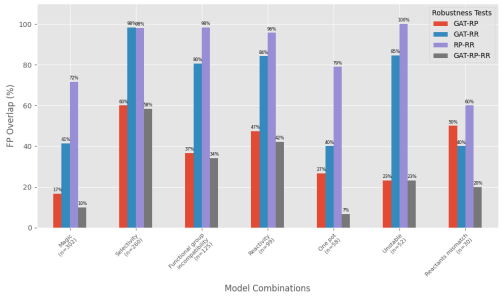

Figure 3: Overlap between individual pairs of scorers and triple of all scorers across different failure categories, with sample sizes indicated for each category. RR and RP show high overlap except in *One-pot*, *Magic*, and *Reactant mismatch* categories. GAT and RP exhibit low overlap across all categories. The joint overlap of all three scorers remains minimal, confirming that each contributes unique discriminative capabilities to the Meta-Scorer.

6.2 Path-Level Plausibility Evaluation

Figure 4 presents the evaluation results comparing our Retro* systems across different configurations: the baseline system (without any scorer), individual scorers (GAT, RR, and RP), and the Meta-Scorer.

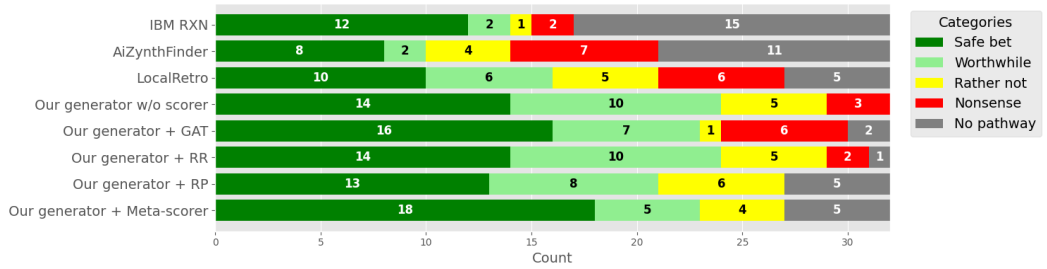

Figure 4: Comparison of our retrosynthesis systems against IBM RXN, AiZynthFinder and LocalRetro. AiZynthFinder, IBM RXN and LocalRetro fail on many targets. Our generator when used without any reaction scorer (indicated as "Our generator w/o scorer") finds pathways for all targets but includes unreliable routes. Introduction of individual scorers trades coverage for reliability, with RP eliminating all *Nonsense* pathways. The search system backed by the Meta-Scorer produces most trustworthy results.

We benchmarked these against two publicly available retrosynthetic planning systems, IBM RXN and AiZynthFinder [8], as well as the open-source generator LocalRetro [3], using the same Retro* search algorithm.

We ran AiZynthFinder in its default configuration from the official repository, with the only modification being an increased time limit of 15 minutes. IBM RXN was used through its free web application [10]. For LocalRetro, we relied on the implementation available in the open-source framework Syntheseus [16]. For all these runs, we used the same starting materials database, eMolecules.

AiZynthFinder demonstrates significant limitations, failing to identify viable pathways for significant number of the target molecules while generating a substantial proportion of unreliable routes classified as *Nonsense* and *Rather Not*. IBM RXN shows improved performance by increasing the number of reliable pathways and reducing hallucinated predictions, yet fails to produce valid synthetic routes for a considerable fraction of target compounds.

Our baseline model (SSR generator without any filtering mechanism) significantly improves number of pathways found, providing solutions for all targets. However, confidence in its results is undermined by the significant presence of unreliable *Nonsenense* and *Rather Not* paths. Introducing individual scorers increases the fraction of targets for which no paths are found, a trade-off that can be desirable for the trustworthiness of the system — rejecting some targets is preferable to mixing reliable and unreliable pathways, as long as the remaining routes are correct. While GAT and RR scorers reduce number of unreliable paths only modestly, our RP scorer demonstrates its value as a primary filter by eliminating all *Nonsense* reactions, though this comes at the cost of fewer *Safe Bet* and *Worthwhile* pathways. Finally, the Meta-Scorer delivers substantial improvements in reliability: significantly increasing *Safe Bet* paths, maintaining zero *Nonsense* results, and reducing *Rather Not* pathways.

7 Conclusions

We introduced a novel framework for mitigating hallucinations in automated retrosynthesis. Unlike previous approaches, our system was evaluated for the first time under a structured validation protocol with Ph.D.-level chemists, focusing on very novel targets. The framework combines a chemist-like primary scorer with two complementary auxiliary scorers, which together form a Meta-Scorer that reliably evaluates the quality of pathways proposed by retrosynthesis models. The scorers exhibit distinct error patterns and, when combined, successfully eliminate chemically unsound reactions. Importantly, our system achieved zero most severe hallucinations, classified as *Nonsense* reactions, in expert evaluation, performing better in this regard than the comparison baselines.

By substantially reducing the number of intermediate-quality reactions (*Rather not* and *Worthwhile*) while increasing the proportion of high-confidence *Safe Bet* reactions, the framework enhances both the efficiency and trustworthiness of retrosynthetic planning. This ability to provide chemists with reliable and validated reaction proposals addresses a central barrier to adoption: the impracticality of assessing large numbers of candidates when unreliable reactions are present. Consequently, our

framework not only improves computational retrosynthesis but also makes it more practical and usable in real-world discovery pipelines.

References

[1] 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.

[2] Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021. doi: 10.1021/jacsau.1c00246. URL https://doi.org/10.1021/jacsau.1c00246. PMID: 34723264.

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

[4] 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.

[5] 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.

[6] 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., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/0d2b2061826a5df3221116a5085a6052-Paper.pdf.

[7] Yannick Djoumbou-Feunang, Jeremy Wilmot, John Kinney, Pritam Chanda, Pulan Yu, Avery Sader, Max Sharifi, Scott Smith, Junjun Ou, Jie Hu, Elizabeth Shipp, Dirk Tomandl, and Siva P. Kumpatla. Cheminformatics and artificial intelligence for accelerating agrochemical discovery. *Frontiers in Chemistry*, Volume 11 - 2023, 2023. ISSN 2296-2646. doi: 10.3389/fchem.2023.1292027. URL https://www.frontiersin.org/journals/chemistry/articles/10.3389/fchem.2023.1292027.

[8] 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, 2020.

[9] Yu Han, Mingjing Deng, Ke Liu, Jia Chen, Yuting Wang, Yu-Ning Xu, and Longyang Dian. Computer-aided synthesis planning (casp) and machine learning: Optimizing chemical reaction conditions. *Chemistry*, page e202401626, 2024. URL https://api.semanticscholar.org/CorpusID:271598676.

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

[11] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, jan 2022. doi: 10.1088/2632-2153/ac3ffb. URL https://dx.doi.org/10.1088/2632-2153/ac3ffb.

[12] 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.

[13] 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, 2019. URL http://arxiv.org/abs/1910.13461.

[14] 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.

[15] Lanxin Long, Rui Li, and Jian Zhang. Artificial intelligence in retrosynthesis prediction and its applications in medicinal chemistry. *Journal of Medicinal Chemistry*, 68(3):2333–2355, 2025. doi: 10.1021/acs.jmedchem.4c02749. URL https://doi.org/10.1021/acs.jmedchem.4c02749. PMID: 39883477.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8818–8827. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/shi20d.html.

[23] 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.

[24] Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10186–10194. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/5470abe68052c72afb19be45bb418d02-Paper.pdf.

[25] Igor V. Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11, 2020. URL https://api.semanticscholar.org/CorpusID:221848923.

[26] 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.

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

[28] Jialu Wu, Yiheng Zhu, Xiaorui Wang, Yitong Li, Mingze Yin, Tianyue Wang, Yuqiang Han, Yu Kang, Yafeng Deng, Jian Wu, Chang-Yu Hsieh, and Tingjun Hou. HiCLR: Knowledge-Induced hierarchical contrastive learning with retrosynthesis prediction yields a reaction foundation model. *JACS Au*, 5(7):3140–3155, June 2025.

[29] Shufang Xie, Rui Yan, Junliang Guo, Yingce Xia, Lijun Wu, and Tao Qin. Retrosynthesis prediction with local template retrieval. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i4.25664. URL https://doi.org/10.1609/aaai.v37i4.25664.

[30] 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.

[31] Chaochao Yan, Peilin Zhao, Chan Lu, Yang Yu, and Junzhou Huang. Retrocomposer: Composing templates for template-based retrosynthesis prediction. *Biomolecules*, 12(9), 2022. ISSN 2218-273X. doi: 10.3390/biom12091325. URL https://www.mdpi.com/2218-273X/12/9/1325.

[32] Lin Yao, Wentao Guo, Zhen Wang, Shang Xiang, Wentan Liu, and Guolin Ke. Node-aligned graph-to-graph: Elevating template-free deep learning approaches in single-step retrosynthesis. *JACS Au*, 4(3):992–1003, 2024. doi: 10.1021/jacsau.3c00737. URL https://doi.org/10.1021/jacsau.3c00737.

[33] Zongao Ye, Limin Yu, and Fei Ma. Semi-template framework for retrosynthesis prediction using graph neural network. *Pattern Recognition*, 168:111825, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111825. URL https://www.sciencedirect.com/science/article/pii/S0031320325004856.

A Examples of reaction pathways

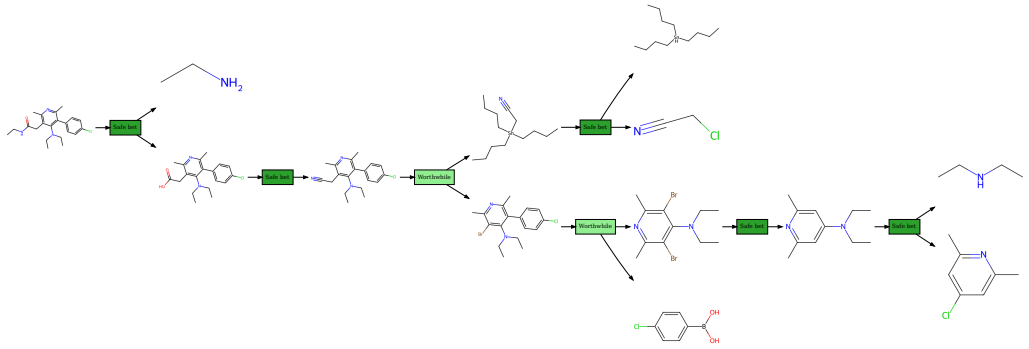

Figure 5: Example of a pathway with Safe Bet and Worthwhile reactions

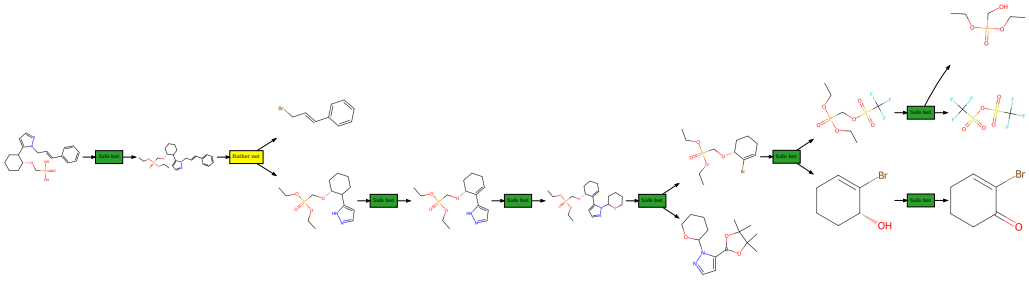

Figure 6: Example of a pathway with a Rather Not reaction

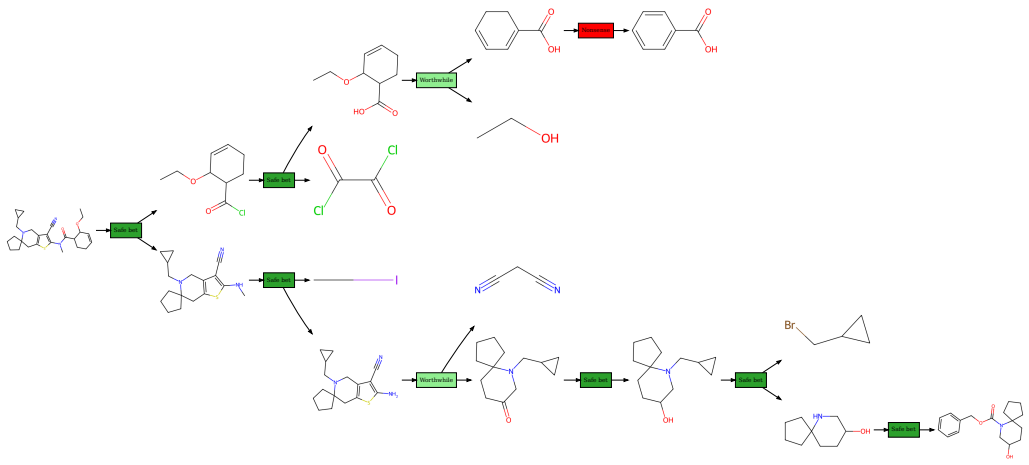

Figure 7: Example of a pathway with a Nonsense reaction

B Reaction Evaluation Protocol

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 inplausibility 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*.

B.1 Implausibility Annotation Examples

B.1.1 Reactants mismatch

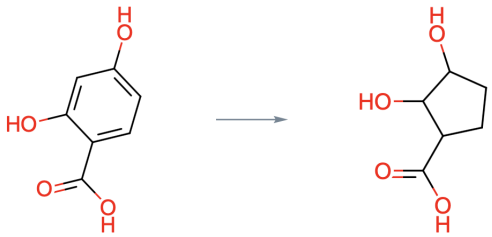

Figure 8: Nonsense: No clear relationship between atoms in the product and the substrate can be confidently proposed

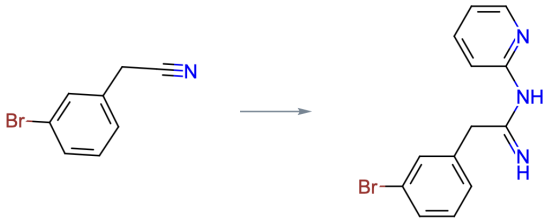

Figure 9: Nonsense: The pyridyl fragment require an additional substrate, that is missing

B.1.2 Unstable

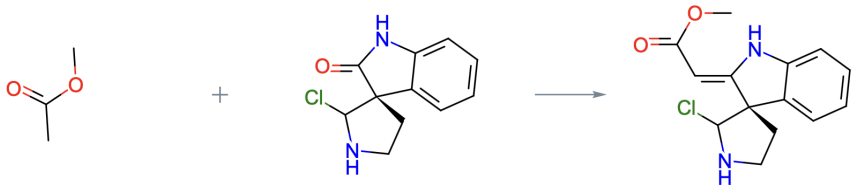

Figure 10: Nonsense: The carbon atom with amine and chlorine is not something seen in literature

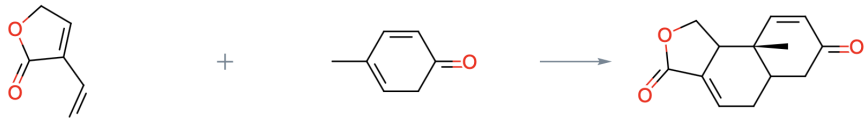

Figure 11: Nonsense: The second substrate would tautomerize to phenol instantly

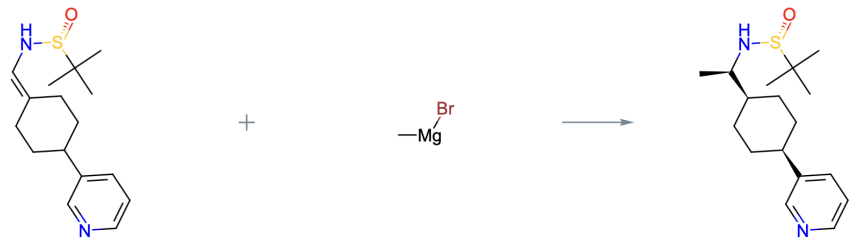

Figure 12: Nonsense: The substrate is unstable, it would tautomerize to imine

B.1.3 Magic

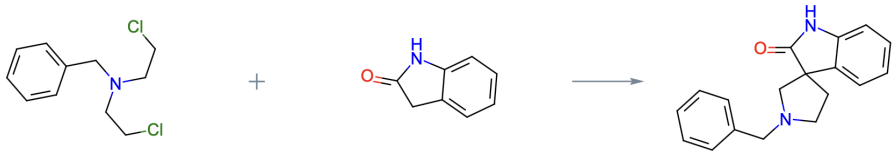

Figure 13: Nonsense: Changing length of the alkyl chain, no known precedent of such variant of carbon alkylation

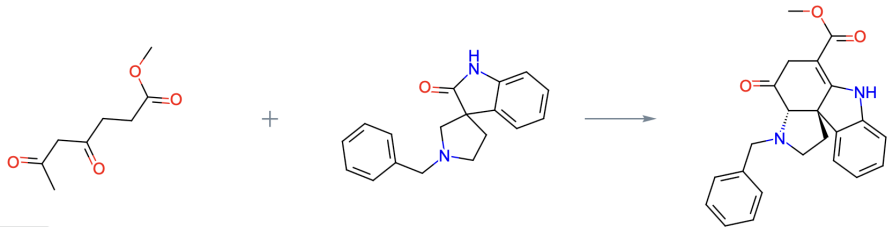

Figure 14: Nonsense: An alkyl chain acting as a leaving group and bond formation by an unactviated amine carbon. No such reactivity ever demonstrated in literature

B.1.4 One pot

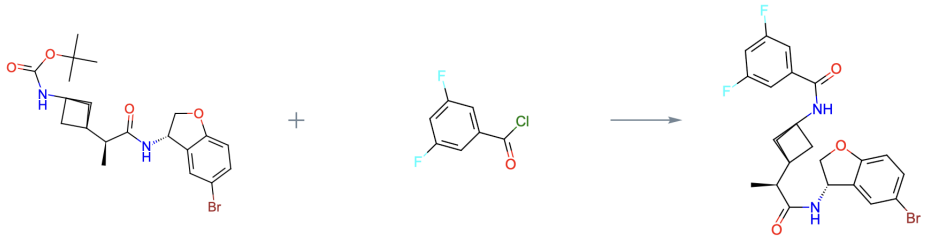

Figure 15: Rather not: 2 steps required – Boc deprotection and acylation

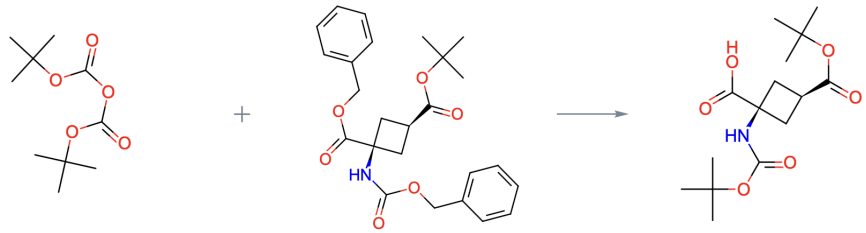

Figure 16: Rather not: 2 steps required - Cbz deprotection and Boc protection

B.1.5 Reactivity

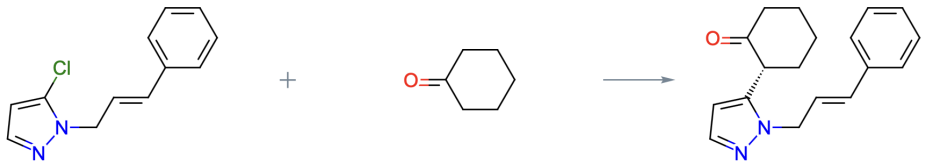

Figure 17: Rather not: Most of the references for this reaction are around electron-deficient heterocycles, only one example with pyrazole in literature

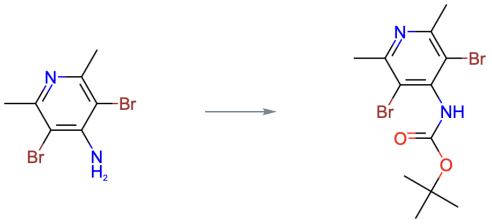

Figure 18: Rather not: High likelihood of steric hindrance

B.1.6 Functional group incompatibility

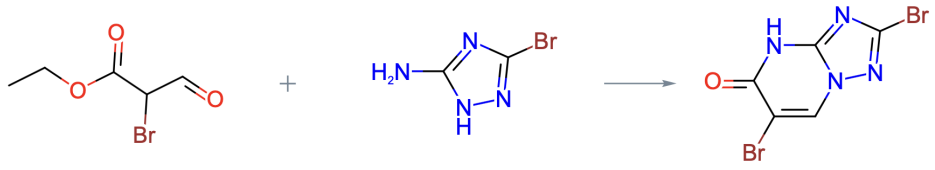

Figure 19: Rather not: No literature references where a bromine is located in alpha to the ester position. The alkyl bromine would most likely react more readily than the ester.

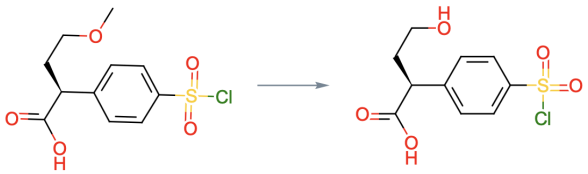

Figure 20: Nonsense: No conditions allow to cleave a methyl ether in a way that wouldn't affect the sulfonyl chloride

B.1.7 Selectivity

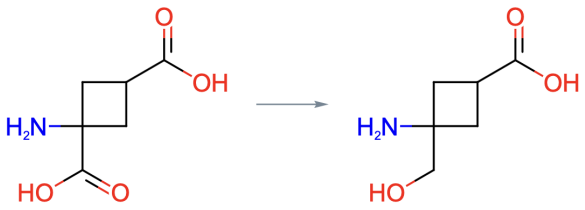

Figure 21: Rather not: There is a considerable risk that achieving the disubstituted product at a satisfactory yield would be very difficult (especially accounting for the presence of amine in the structure).

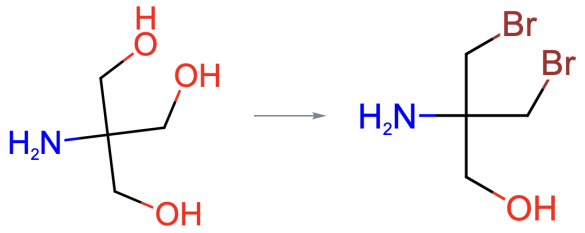

Figure 22: Rather not: There are 3 equivalent hydroxyl groups, so in bromination we expect triple substitution rather than this scenario

C Retrosynthesis Targets

C.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

C.2 Visualization

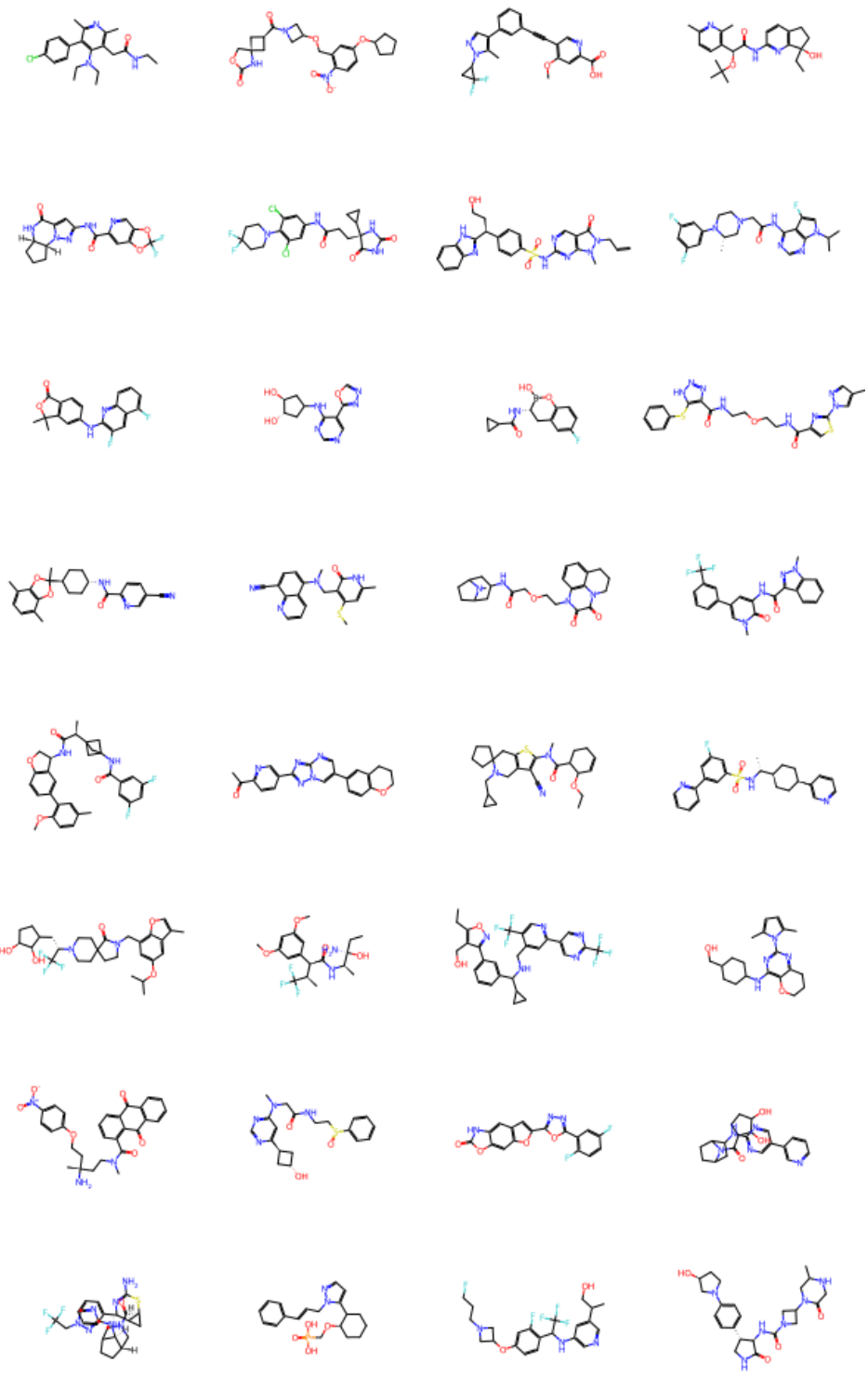

Figure 23: 32 molecules that have been used as targets for retrosynthesis.

D ROC And Precision-Recall Curves By Failure Category

D.1 Magic

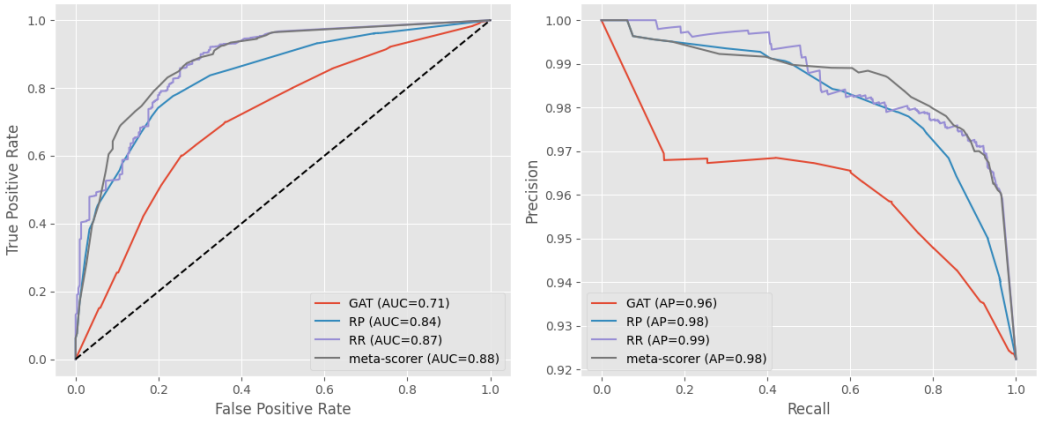

Figure 24: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Magic* and *No Problem* reactions.

D.2 Selectivity

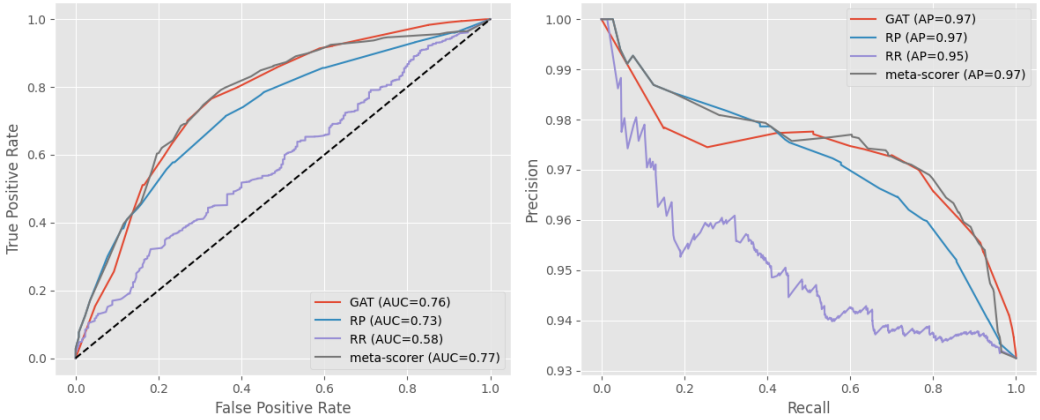

Figure 25: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Selectivity* and *No Problem* reactions.

D.3 Functional group incompatibility

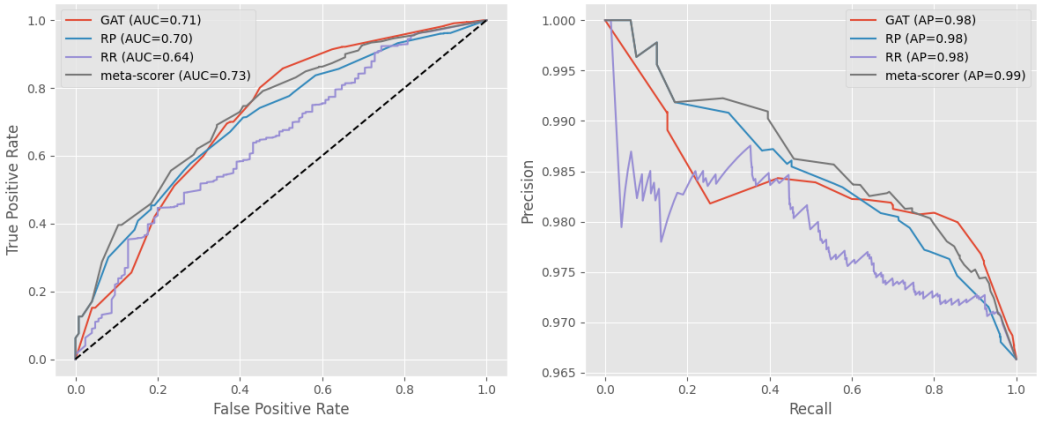

Figure 26: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Functional group incompatibility* and *No Problem* reactions.

D.4 Reactivity

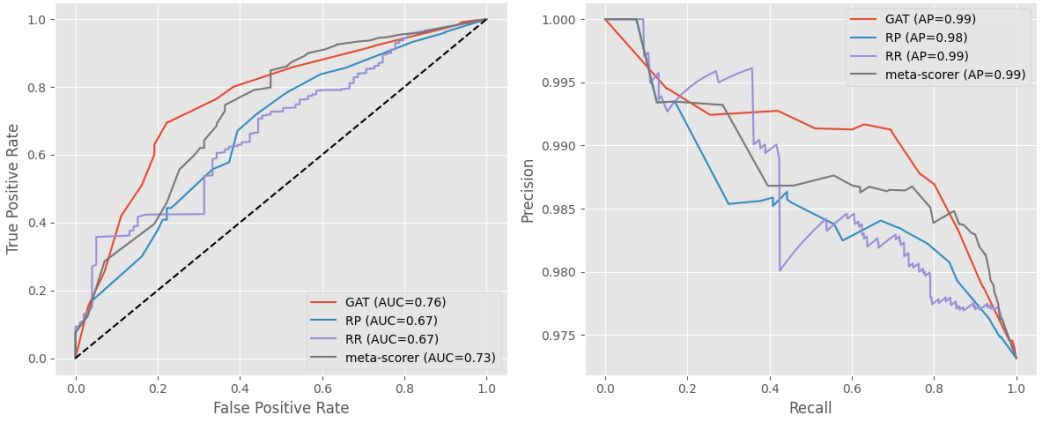

Figure 27: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Reactivity* and *No Problem* reactions.

D.5 One pot

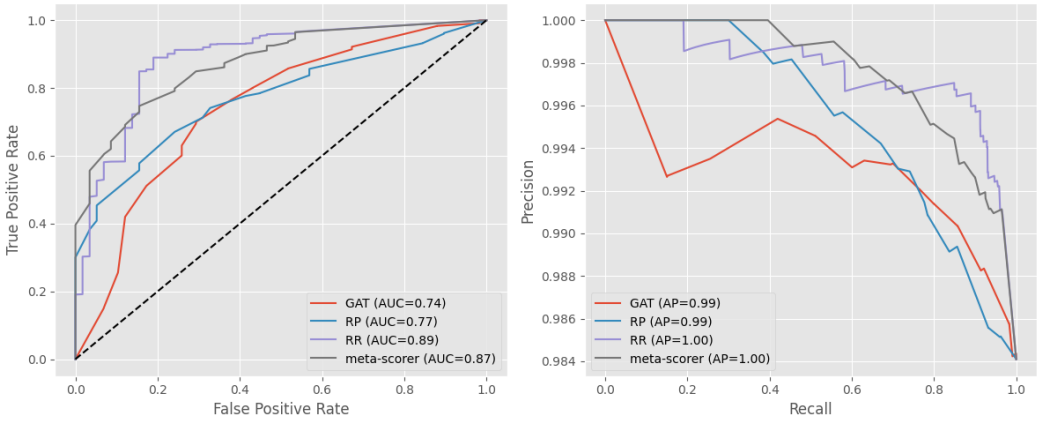

Figure 28: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *One pot* and *No Problem* reactions.

D.6 Unstable

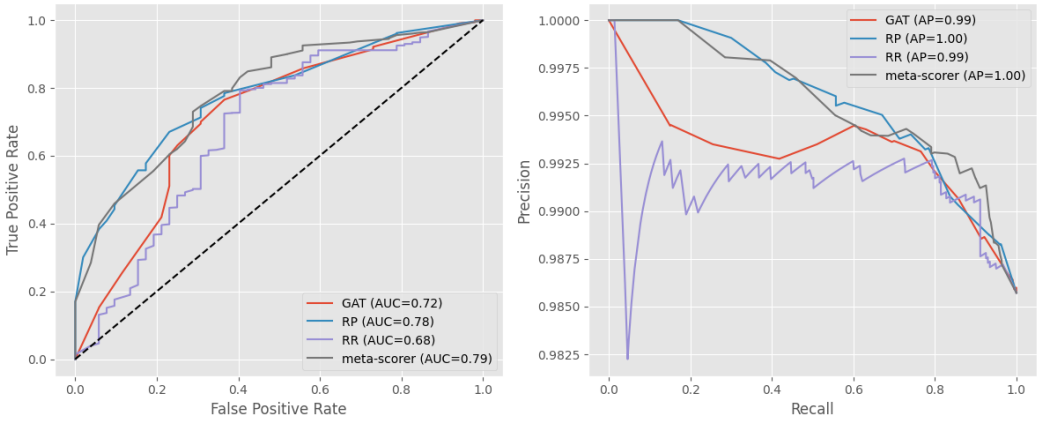

Figure 29: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Unstable* and *No Problem* reactions.

D.7 Reactants mismatch

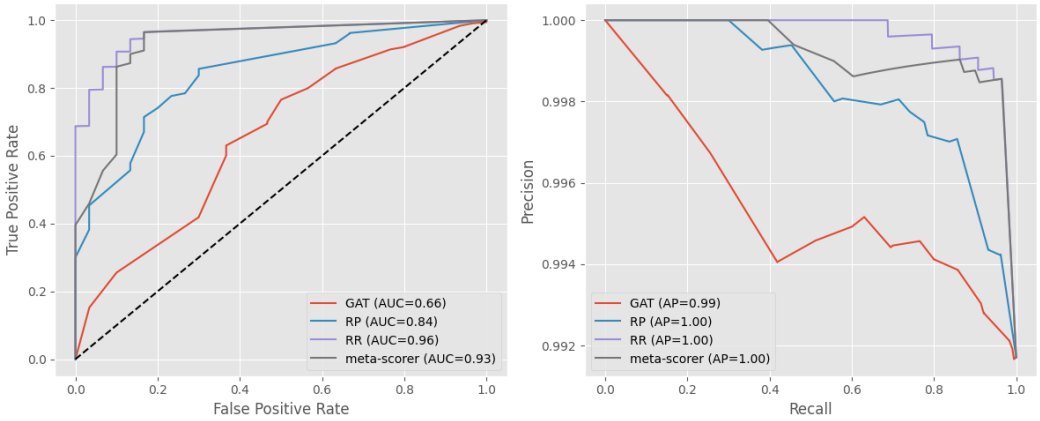

Figure 30: ROC (on the left) and precision-recall (on the right) curves comparing the performance of individual scorers versus the Meta-Scorer on *Reactants mismatch* and *No Problem* reactions.

E False Positives Counts By Failure Category

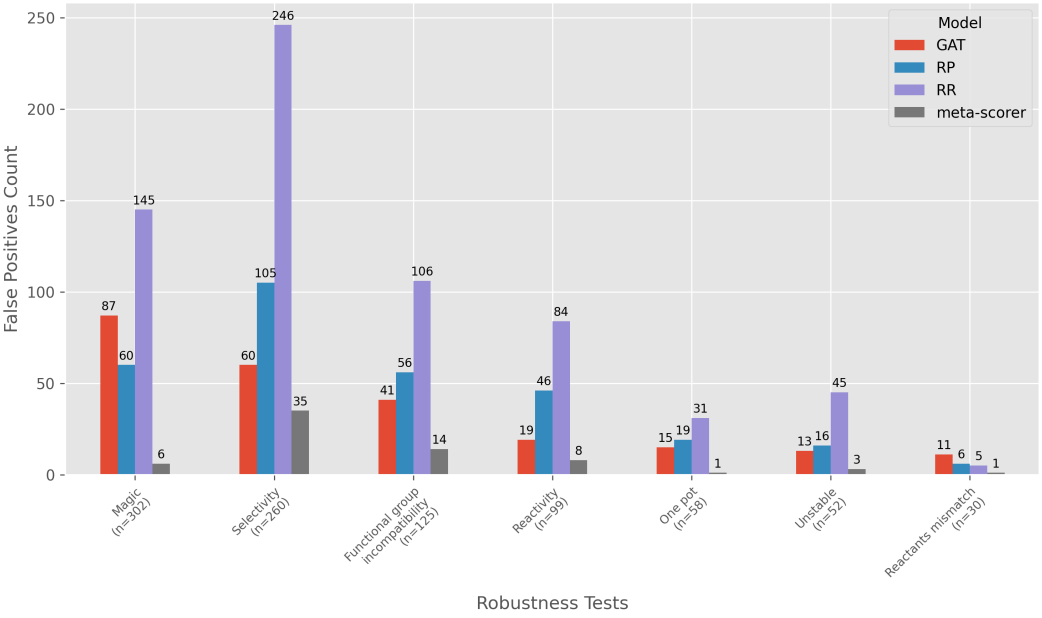

Figure 31: Counts of false positives produced by individual scorers versus the Meta-Scorer across different failure categories, with sample sizes indicated for each category.

F True Negatives Counts By Failure Category

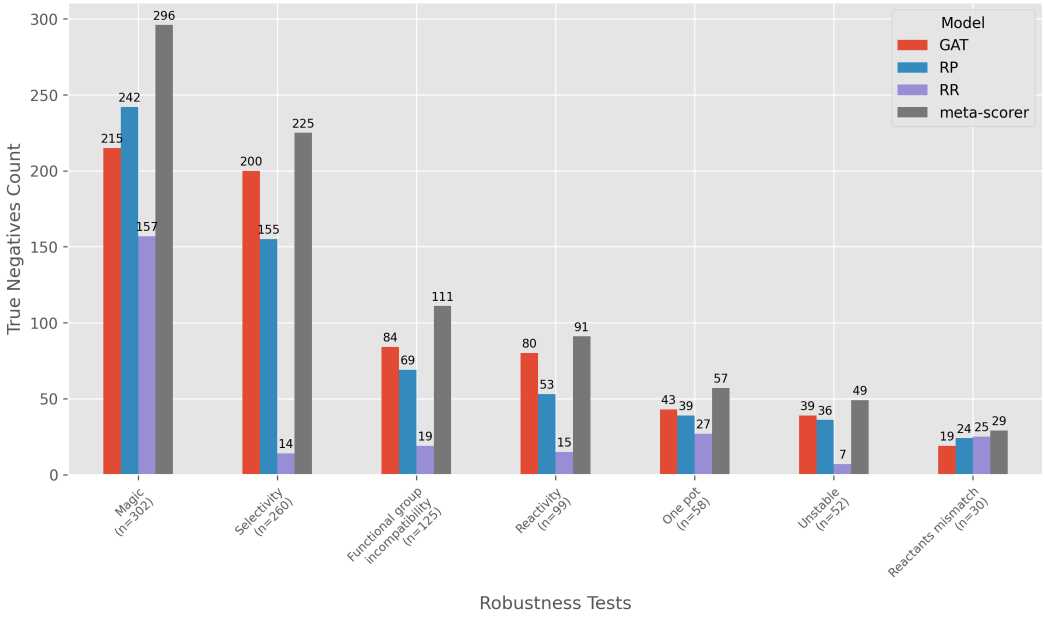

Figure 32: Counts of true negatives produced by individual scorers versus the Meta-Scorer across different failure categories, with sample sizes indicated for each category.