RETROROUTE: ROUTE-LEVEL AGENTIC DECISION MAKING FOR MULTI-STEP RETROSYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Development of robust and effective strategies for retrosynthetic planning requires effective decision making throughout of the synthesis process. A critical step in achieving this goal is enabling an agent to evaluate intermediate choices according to their long-term contribution to successful route completion. Current machine learning-based methods often focus on single-step prediction or heuristic search, optimizing locally plausible reactions without explicitly learning how sequential decisions affect the final synthetic route, resulting in suboptimal planning for long and complex syntheses. Therefore, we introduce RetroRoute, an advanced agentic framework for multi-step retrosynthesis that optimizes decisions over the entire synthesis route. The proposed LLM-based REVIEWER and state-aware ANALYZER form the RetroRoute agent for candidate generation and route-level decision making. First, the planning-conditioned REVIEWER generates and reviews candidate reactions based on the target, current intermediate, and search budget, with route-aware utility refinement further identifying chemically feasible and route-compatible alternatives. Furthermore, the state-aware ANALYZER represents each search state with search history and learns to rank candidate reactions toward successful route completion. As a result, RetroRoute optimizes sequential decisions for long-term route utility, enabling agentic global planning toward synthesis completion. In experiments on the USPTO pathways dataset RetroBench, RetroRoute outperforms state-of-the-art methods, achieving up to a 3% improvement in Top-1 test accuracy, particularly for long synthetic routes. These results demonstrate the superiority of RetroRoute in using route-level information to guide agentic decisions over the complete retrosynthetic planning process. They also demonstrate its potential for advancing autonomous retrosynthetic planning and facilitating reliable synthesis route design. Code is available at https://github.com/Su-yq/RetroRoute.

1 INTRODUCTION

Retrosynthetic planning aims to identify a sequence of chemical transformations that decomposes a target molecule into available starting materials. It is a central problem in computer-aided synthesis planning and has applications in drug discovery, materials design, and automated chemistry (Chen et al., b; Lee et al., 2026; Tu et al., 2025). Modern systems commonly combine a single-step reaction model (Xuan-Vu et al., 2025; Zheng et al., 2019; Yan et al., 2020) with a search procedure that recursively expands candidate disconnections until a complete route reaches purchasable starting materials (Segler et al., 2018; Chen et al., 2020).

This decomposition has driven substantial progress, but it also creates an objective mismatch (Torren-Peraire et al., 2024). Proposal models are typically trained on individual reaction pairs to predict plausible reactants for a given product (Maziarz et al., 2023), whereas planning is evaluated at the route level, where each action must also support downstream synthesizability and remain compatible with the evolving search state and budget (Liu et al., 2025b). Consequently, a locally plausible reaction with a high proposal score may lead to a suboptimal search state for subsequent planning (Long et al., 2025; Liu et al., 2025a). This mismatch becomes increasingly important for longer synthesis routes, where the effects of successive local decisions accumulate (Shee et al., 2025; Mishra et al., 2026; Bran et al.).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent planning methods increasingly incorporate route-level information beyond local reaction likelihood. Learned value functions, route context, preference models, uncertainty estimation, and constrained search have all improved multistep decision making (Chen et al., 2020; Liu et al., 2023; 2024a; Tripp et al., 2024). In parallel, chemistry specific language modeling has shown that domain adaptation and reaction aware representations can improve chemical information extraction and retrosynthetic prediction (Zhang et al., 2024; Xiong et al., 2025). Large language Models (LLMs) have further been introduced as route-level reasoning and decision components, including context aware prediction, agent based planning, and LLM-guided search (Kang et al., 2025; Liu et al., 2025d; Wang et al., 2026; Song et al., 2026). These developments motivate a more explicit connection between reaction proposal, training supervision, and state dependent route decisions.

Building on these studies, we adopt an agentic decision-making formulation for retrosynthetic planning, where each action is evaluated according to its long-term contribution to route completion. This formulation further addresses the inherent multimodality of retrosynthesis, where multiple valid disconnections may lead to the same target molecule (Zagribelnyy et al., 2026). A reaction database usually records a small number of synthetic choices for a product, while other chemically plausible disconnections may also be useful (Pappala et al., 2026). Generated candidates that differ from the recorded reaction therefore provide uncertain negative evidence. Aggressively suppressing all such candidates can narrow the proposal distribution in a way that is poorly matched to route search (Hassen et al., 2022). Instead, retrosynthetic planning requires distinguishing useful alternative reactions from candidates that are unlikely to support successful route completion (Chen et al., a). This motivates evaluating generated actions using both route-level feasibility and their compatibility with the overall synthesis route (Yuan et al., 2024; Tripp et al., 2024).

To address these limitations, we develop RetroRoute, an agentic framework that reformulates multistep retrosynthesis from local reaction prediction into route-level sequential decision making. Rather than optimizing each reaction independently, RetroRoute evaluates intermediate actions by their long-term utility for completing the entire synthesis route. This formulation is realized through three coupled stages. First, we propose a Reviewer, instantiated with MolT5, that incorporates the target, current intermediate, search depth, and search budget to generate reactions under the current planning state. Second, generated alternatives are evaluated using chemical feasibility and route utility, extending supervision beyond a single reference reaction. Third, an Analyzer, instantiated with ChemDFM(Zhao et al., 2025), learns a state-aware decision policy over the evolving search state to rank candidate reactions according to their expected contribution to route completion. The selected actions continuously update the search state and guide subsequent decisions, establishing a closed-loop interaction between reaction generation, route evaluation, and sequential decision making. In this way, RetroRoute provides a unified framework for optimizing local reaction choices toward long-term retrosynthetic objectives while preserving an explicit and traceable route construction process. Our main contributions are summarized as follows:

- We propose RetroRoute for multi-step retrosynthesis as route-level agentic decision making, shifting the optimization objective from isolated reaction accuracy to long-term route completion.
- We develop a planning scaffolded Reviewer for reaction candidate generation with utility guided policy refinement, enabling context aware candidate generation and extending supervision beyond the reference to chemically and route supported alternatives.
- We introduce a state-aware Analyzer that jointly evaluates candidate reactions over the evolving search state and optimizes sequential decisions toward successful route completion.

2 RELATED WORK

Single Step Retrosynthesis. Single-step retrosynthesis defines the local reaction action space used by most multistep planners(Zheng et al., 2022). Early approaches rely on reaction templates and template ranking (Coley et al., 2017; Segler & Waller, 2017; Dai et al., 2019), while graph based methods identify reaction centers and reconstruct reactants through molecular transformations (Liu et al., 2024b; Shi et al., 2020; Somnath et al., 2021; Sacha et al., 2021). Template free sequence models instead formulate retrosynthesis as direct product to reactant generation (Liu et al., 2017; Edwards et al., 2022). More recent methods incorporate explicit chemical reasoning and verifiable

feedback into reaction prediction. RetroDFM R combines supervised reasoning with chemically grounded reinforcement signals (Zhang et al., 2026), while RetroReasoner learns structured disconnection rationales with round trip consistency feedback (Ko et al., 2026). These methods primarily improve local reaction quality. RetroRoute complements this direction by aligning proposal generation with the downstream planning state and by refining the proposal policy using route informed utility.

Multistep Retrosynthetic Planning. Multistep retrosynthesis is commonly formulated as sequential decision making over a molecular search space. Monte Carlo Tree Search with learned reaction policies established an early neural planning paradigm (Segler et al., 2018), and Retro* introduced learned value estimation within an AND-OR search framework (Chen et al., 2020). Subsequent work incorporates additional route-level information through goal conditioned policies, route context, preference learning, uncertainty estimation, and constrained planning (Yu et al., 2022; Liu et al., 2023; 2024a; Tripp et al., 2024; Yuan et al., 2024; Yu et al., 2024). Reinforcement learning has also been used to optimize route policies beyond supervised reaction likelihood (Guo et al., 2024; Wang & Montana, 2025), while RetroSynFormer models synthesis trajectories using reward conditioned sequential decision making (Granqvist et al., 2026). RetroRoute differs in explicitly separating candidate coverage from state dependent branch evaluation while connecting the two through utility guided proposal refinement.

LLM-Guided Retrosynthesis. LLMs increasingly serve as chemical reasoning and route decision components. RetroInText incorporates textual route context into multistep retrosynthetic prediction (Kang et al., 2025), while LLM augmented synthesis decision programs use compact pathway representations to guide search (Wang et al., 2025). Agent based approaches such as Retro R1 train LLM policies to interact with retrosynthesis tools and environmental feedback (Liu et al., 2025d). Other methods study direct route reasoning, reinforcement learning, explicit search, and memory based planning (Wang et al., 2026; Song et al., 2026; Zuo et al., 2026; Zhu et al., 2026). These studies demonstrate the value of global route context for retrosynthetic reasoning. RetroRoute uses the chemistry LLM in a narrower decision role, where it evaluates a fixed candidate set under the evolving search state while route construction remains explicit.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let $\mathcal{M}$ denote the molecular space and $\mathcal{B} \subset \mathcal{M}$ the set of purchasable starting materials. Given a target molecule $M_0 \in \mathcal{M}$, retrosynthetic planning aims to recursively decompose M_0 through a sequence of retrosynthetic actions until all terminal molecules belong to $\mathcal{B}$. At search step t, we represent the planning state as:

$$s_t = (M_0, M_t, \mathcal{F}_t, \mathcal{H}_t, d_t, D)\,, \tag{1}$$

where M_t is the molecule selected for expansion, $\mathcal{F}_t$ is the unresolved frontier, $\mathcal{H}_t$ records previous retrosynthetic decisions, and d_t and D denote the current and maximum search depths, respectively. And conditioned on s_t, a proposal policy generates a set of K candidate retrosynthetic actions by Equation 2:

$$\mathcal{C}_t = \left\{c_t^1, \ldots, c_t^K\right\}, \tag{2}$$

where the search controller selects an action $c_t \in \mathcal{C}_t$, which expands the corresponding branch and updates $\mathcal{F}_t$ and $\mathcal{H}_t$, resulting in the next state s_{t+1}. This process continues until all terminal molecules are purchasable or the search limit is reached.

RetroRoute realizes this sequential decision process through a route-aware agentic framework, as illustrated in Figure 1. At each search step, the Reviewer generates chemically plausible candidate reactions conditioned on the current planning state. We further introduce Utility-Regularized Positive Optimization (URPO) to refine the proposal policy toward candidates that are both chemically feasible and beneficial for downstream route completion. Given these candidates, the chemistry LM serves as a state-aware decision agent that incorporates the search history to rank candidate reactions according to their long-term route utility. Together, these components enable RetroRoute to move beyond locally plausible reaction prediction toward route-level decision making for multi-step retrosynthetic planning.

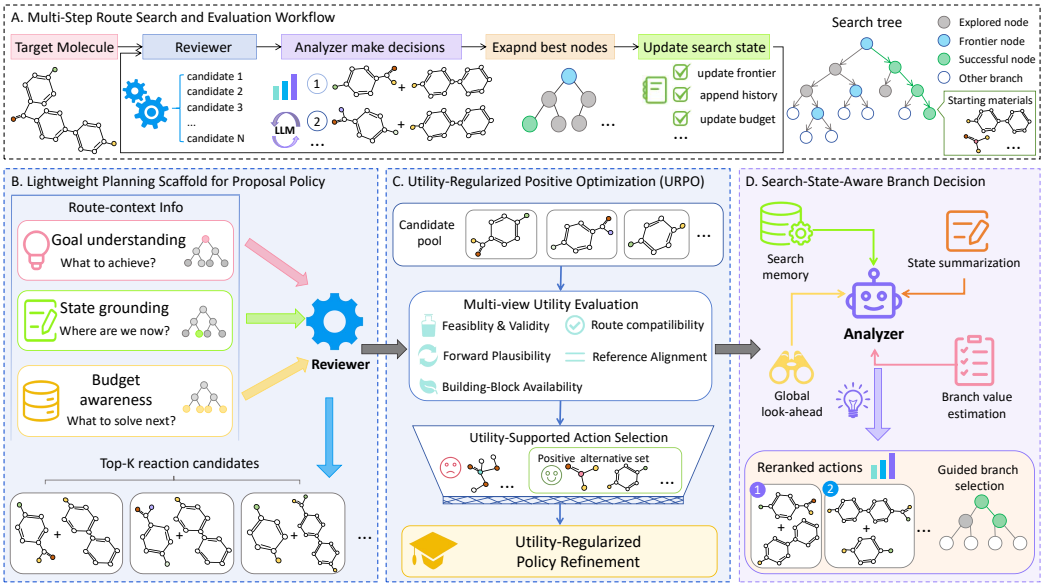


Figure 1: **Overview of RetroRoute.** **A** Closed-loop route planning integrating proposal, decision, expansion, and state update. **B** A planning scaffold guides the Reviewer with goal, state, and budget information. **C** URPO identifies utility-supported reaction actions and refines the proposal policy. **D** The Analyzer performs state-aware listwise branch selection during search.

3.2 Planning-Scaffolded Proposal Policy

Single-step retrosynthesis models typically estimate reaction likelihood from the molecule currently being decomposed. Such local conditioning provides limited information about the planning process in which the reaction will eventually be used. We therefore expose a compact planning scaffold to the Reviewer. For state s_t, the proposal context is $x_t = [M_0, M_t, d_t, D, g_{\mathcal{B}}]$, where $g_{\mathcal{B}}$ specifies the synthesis objective of reaching purchasable starting materials. And the Reviewer defines the policy as $\pi_\theta(c_t \mid x_t)$.

The scaffold separates target information, local reaction state, and search budget while keeping the prediction space unchanged. The model still outputs only a reactant sequence, allowing the reaction prior to incorporate planning-relevant context without requiring an explicit reasoning trace.

All scaffold variables are observable during search. Future intermediates, reference reactants, route identifiers, and ground-truth remaining depth are excluded from x_t. Beam decoding from π_θ produces $\mathcal{C}_t$, which forms both the action space explored by URPO during training and the reaction space evaluated during multistep planning.

3.3 Utility Regularized Positive Optimization

A recorded reaction provides one demonstrated disconnection for a product, while multiple alternative transformations may remain chemically feasible. Consequently, treating every nonreference generation as negative evidence can suppress useful regions of the proposal distribution. This ambiguity becomes particularly important in multistep planning, where different local disconnections may lead to distinct viable synthesis routes.

URPO performs utility-guided offline refinement of the proposal policy. For each training context x, we sample a candidate set $\mathcal{C}(x)$ from π_θ and evaluate each candidate using complementary chemistry- and route-level signals that quantify its value for continued policy optimization. We integrate these signals into the following utility function:

$$U(c;x) = \lambda_{\text{chem}} u_{\text{chem}}(c) + \lambda_{\text{fwd}} u_{\text{fwd}}(c) + \lambda_{\text{route}} u_{\text{route}}(c) + \lambda_{\text{ref}} u_{\text{ref}}(c) + \lambda_{\text{stock}} u_{\text{stock}}(c) - \lambda_{\text{bad}} P_{\text{bad}}(c), \tag{3}$$

where u_{chem} measures molecular validity, u_{fwd} evaluates whether an independent forward reaction model can reconstruct the current product (Schwaller et al., 2019), and u_{route} measures compatibility with the synthesis trajectory available during training. The reference term u_{ref} measures structural and representational consistency with the demonstrated transformation, while u_{stock} rewards progress toward purchasable starting materials. The penalty P_{bad} suppresses undesirable reaction actions.

Since the aggregate utility may allow a high score on one signal to compensate for a low score on another, URPO further imposes admissibility constraints when selecting candidate actions:

$$A(c;x) = \prod_{r=1}^{R} \mathbb{I}\left[q_r(c;x) \in \mathcal{I}_r\right], \tag{4}$$

$$\mathcal{P}_U(x) = \text{Top}\left\{c \in \mathcal{C}(x) \mid A(c;x) = 1,\ U(c;x) \geq \tau\right\}, \tag{5}$$

here, $q_r(c;x)$ denotes the r-th candidate quality signal and $\mathcal{I}_r$ its admissible region. Thus, $A(c;x) = 1$ only when all specified quality signals satisfy their corresponding admissibility conditions. Among these candidates, $\mathcal{P}_U(x)$ retains those with utility above the threshold τ, with $\text{Top}(\cdot)$ selecting candidates according to their utility scores. Thus, candidate selection considers both admissibility and aggregate utility.

This construction defines a constrained high-utility region within the action support explored by the current proposal policy. Per-context and global support controls further regulate the contribution of self-generated actions relative to demonstrated supervision.

The supported actions are distilled back into the proposal policy together with the demonstrated reactions as Equation 6, and the resulting policy update as Equation 7:

$$\mathcal{D}_{\text{URPO}} = \mathcal{D}_{\text{gold}} \cup \left\{(x,c) \mid c \in \mathcal{P}_U(x)\right\}, \tag{6}$$

$$\mathcal{L}_{\text{URPO}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{URPO}}} \log \pi_\theta(y \mid x). \tag{7}$$

Therefore, URPO performs conservative offline policy refinement over self-generated reaction actions. Utility serves as an offline reward surrogate that determines which regions of the proposal policy receive additional support, while the policy update preserves a stable positive-likelihood objective. Since rejected sequences are not required for optimization, chemically ambiguous alternatives do not need to be assigned explicit negative labels.

Route-level supervision derived from synthesis annotations is available only during training. These signals are used exclusively for offline action evaluation and are excluded from both the proposal input and the test-time search state. Consequently, URPO leverages route-level information for policy refinement without altering the inference interface of the proposal policy.

3.4 Search-State-Aware Branch Decision Model

The proposal policy provides candidate coverage, but local reaction likelihood alone does not determine the value of an action within a partially constructed route. The same disconnection may have different planning value depending on unresolved molecules, previous decisions, available stock, and the remaining search budget. RetroRoute therefore employs a state-aware Analyzer for branch decision making.

For each expansion, the state-aware Analyzer jointly observes the target molecule, current molecule, frontier, route history, remaining depth budget, stock status, and the complete candidate set. It assigns as Equation 8, followed by Equation 9:

$$z_i = f_\phi\left(c_t^i, s_t, \mathcal{C}_t\right), \tag{8}$$

$$p_\phi\left(c_t^i \mid s_t, \mathcal{C}_t\right) = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)}. \tag{9}$$

The frontier represents unresolved synthesis obligations, while the route history provides compact memory of earlier decisions. Joint presentation of the complete candidate set allows the LLM to

compare alternative reactions within a common planning context and assign state-dependent branch values.

Let $\mathcal{P}_t$ denote the supported actions for state s_t. We optimize the listwise (Cao et al., 2007; Liu et al., 2025c)objective:

$$\mathcal{L}_{\text{list}} = -\log \frac{\sum_{i \in \mathcal{P}_t} \exp(z_i)}{\sum_{j \in \mathcal{C}_t} \exp(z_j)}. \tag{10}$$

The objective assigns probability mass to the complete positive subset and naturally accommodates multiple viable reactions within one state. Candidate identifiers are randomized during training to reduce positional dependence. The Analyzer is adapted through Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2022), preserving its pretrained chemical knowledge while learning a route-level decision interface.

3.5 Closed-Loop Multistep Planning

RetroRoute integrates the Reviewer and Analyzer within iterative beam search. Starting from M_0, an unresolved molecule is selected from the frontier and the proposal policy constructs its reaction candidate set. The Analyzer evaluates these actions under the complete current search state, after which selected reactions are expanded into successor states.

Expansion replaces M_t with the proposed reactants and updates $\mathcal{F}_t$, $\mathcal{H}_t$, and d_t. Reactants found in $\mathcal{B}$ are marked terminal. A route is solved when every molecule in the frontier is purchasable.

For an action sequence $\mathcal{A} = \{a_1, \ldots, a_T\}$, the route score accumulates state-conditioned decision probabilities:

$$S(\mathcal{A}) = \sum_{t=1}^{T} \log p_\phi \left(a_t \mid s_t, \mathcal{C}_t\right). \tag{11}$$

Equivalent successor states are deduplicated, and the highest-scoring states are retained in the route beam. Each selected action changes the frontier and route history observed by subsequent LLM decisions. The resulting feedback loop couples reaction proposal with route-level reasoning throughout the entire search process.

4 Experiments

Datasets. We evaluate RetroRoute on RetroBench (Liu et al., 2023), a multistep retrosynthetic planning benchmark constructed from the USPTO reaction network. It contains $46,458$ training targets, $5,803$ validation targets, and $5,838$ test targets with reference synthesis routes. Single step reactions extracted from the training routes are used to adapt the Reviewer. All RetroRoute variants use the same train, validation, and test splits. The purchasable starting material stock $\mathcal{B}$ is derived from the ZINC database (Irwin et al., 2012), following the RetroBench protocol. Route annotations used by URPO are restricted to training and are never exposed during test time planning. Test routes span ground truth depths from 2 to 13. Depth specific results for the primary groups are reported in the main analysis, while the sparse deepest groups are summarized in Appendix C.

Baselines. We compare RetroRoute with representative methods under the Retro* framework. RetroSim is template-based, retrieving templates by molecular similarity (Coley et al., 2017). G2Gs (Shi et al., 2020) and GraphRetro (Somnath et al., 2021) use graph transformations to predict reaction centers and structural edits. MEGAN models reactions as sequences of graph edits (Sacha et al., 2021), while Transformer directly generates reactant sequences in a template-free manner (Karpov et al., 2019). We also include GraphRetro+CREBM and Transformer+CREBM, where CREBM refines candidate evaluation with learned preference scores (Liu et al., 2024a). These baselines cover template-based, graph-based, and sequence-based models with learned route preference refinement.

Evaluation Metrics. Following the RetroBench protocol, we report Top-k route exact match accuracy for $k \in \{1, 2, 3, 4, 5\}$. A predicted route is considered correct when its terminal starting materials match at least one reference set for the same target. Molecular identity is determined using the connectivity block of the InChIKey (Liu et al., 2023). We additionally evaluate Greedy

DFS to examine branch selection when only one local action is retained at each expansion. Model invocation counts are reported as supplementary search diagnostics.

Implementation Overview. We instantiate the Reviewer with Molecular T5 (MolT5) (Edwards et al., 2022) and the state-aware Analyzer with Dialogue Foundation Model for Chemistry (ChemDFM) (Zhao et al., 2025). The Reviewer is initialized from the MolT5 checkpoint used by RetroInText (Kang et al., 2025) and is subsequently adapted on single step reactions extracted from the RetroBench training routes. During multistep search, the Reviewer generates 10 reaction candidates, 5 actions are retained per expansion, and the route beam size is 5. The maximum search depth and number of search rounds are both 14. All ablations use the same test set, starting material stock, and search budget. Full prompt formats, training settings, URPO construction criteria, LLM adaptation, and search parameters are provided in Appendix A.

5 RESULTS AND ANALYSIS

5.1 MAIN RESULTS

The performance of all methods is presented in Table 1. Compared with the baseline methods, our model RetroRoute achieves the best performance, exceeding the Top-1 exact match accuracy of RetroSim by 2.5% and Transformer with CREBM by 2.6%, achieving State-Of-The-Art (SOTA) performance. RetroRoute also demonstrates competitive performance across different Top-k settings, highlighting its effectiveness in multi-step retrosynthetic planning. These results demonstrate the benefits of planning-conditioned reaction generation, URPO, and state-aware decision making. Additional single-step proposal results are provided in Appendix B, further examining how planning-conditioned generation and URPO affect reaction candidate quality before multistep search.

Method	Top-1	Top-2	Top-3	Top-4	Top-5
G2Gs (Shi et al., 2020)	5.4	8.3	9.9	10.9	11.7
GraphRetro (Somnath et al., 2021)	15.3	19.5	21.0	21.9	22.4
GraphRetro + CREBM (Liu et al., 2024a)	16.3	20.1	21.6	22.3	22.7
MEGAN (Sacha et al., 2021)	18.8	27.9	32.7	36.6	38.1
Transformer (Karpov et al., 2019)	31.3	40.4	44.7	47.2	48.9
Transformer + CREBM (Liu et al., 2024a)	35.0	43.4	46.7	48.7	49.7
RetroSim (Coley et al., 2017)	35.1	40.5	42.9	44.0	44.4
RetroRoute	**37.6**	**43.7**	**46.7**	**48.9**	**50.1**

Table 1: Summary of retrosynthetic planning results for exact match accuracy (%) on RetroBench with Retro*.

5.2 COMPARISON UNDER GREEDY DFS

To further evaluate the planning capability of RetroRoute without relying on complex search algorithms, we compare it with SOTA retrosynthesis models under Greedy DFS. As shown in Table 2, RetroRoute achieves the best performance, exceeding the Top-1 exact-match accuracy of the SOTA method FusionRetro by 1.5%. The consistent improvement under Greedy DFS further demonstrates the effectiveness of RetroRoute in reaction selection and multi-step retrosynthetic planning. Detailed Greedy DFS results across route depths, together with search success diagnostics, are reported in Appendix D.

5.3 ABLATION ANALYSIS ACROSS ROUTE DEPTHS

To better evaluate the performance of RetroRoute across different levels of retrosynthetic complexity, we analyze the exact-match accuracy at different route depths under beam search and Greedy DFS, as shown in Figure 2. We focus on route depths from 2 to 6, which cover the majority of targets in the test set.

Under beam search (Figure 2A), RetroRoute achieves the best performance across all evaluated route depths. Compared with the URPO proposal model, RetroRoute improves the accuracy by ap-

Method	Greedy DFS Top-1
GraphRetro (Somnath et al., 2021)	14.4
Transformer (Karpov et al., 2019)	26.7
RetroSim (Coley et al., 2017)	31.5
MEGAN (Sacha et al., 2021)	32.9
FusionRetro + CREBM (Liu et al., 2024a)	33.8
FusionRetro (Liu et al., 2023)	35.2
RetroRoute	**36.7**

Table 2: Greedy DFS Top-1 exact match accuracy (%) on RetroBench.

proximately 9.6%, 11.0%, and 8.4% at depths 4, 5, and 6, respectively. RetroRoute also maintains a more stable performance as route depth increases, while the other variants show a larger performance drop. These results demonstrate the benefit of combining planning-conditioned generation, URPO, and state-aware decision making, particularly for long and complex synthesis routes.

A similar trend is observed under Greedy DFS (Figure 2B). RetroRoute consistently outperforms the proposal-only variants across different route depths, with clear improvements at depths 5 and 6. In addition, Base ChemDFM performs substantially worse than RetroRoute when using the same candidate reactions, showing that pretrained chemical knowledge alone is insufficient for effective reaction selection. The consistent improvements under both beam search and Greedy DFS demonstrate the importance of state-aware decision making for multi-step retrosynthetic planning. Complete Top-1 and Top-5 results across the full range of route depths are provided in Appendix C.

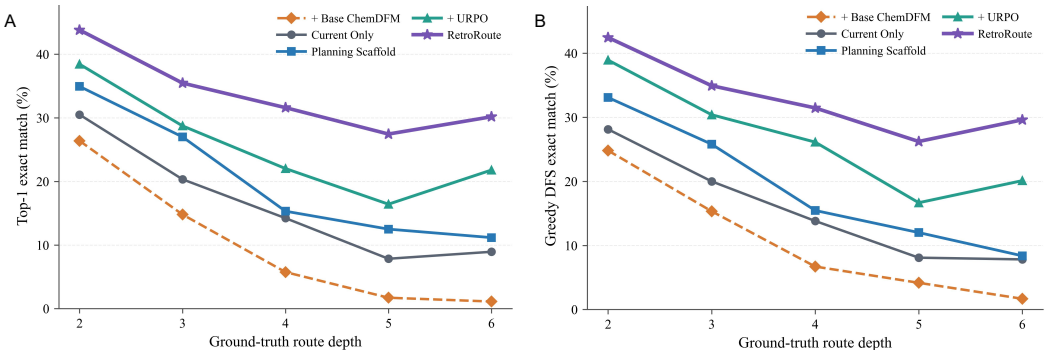


Figure 2: Top-1 route exact-match accuracy across ground-truth route depths under **A** beam search and **B** Greedy DFS.

5.4 Case Study

To further analyze the effect of state-aware decision making, we present a representative example in Figure 3, where the base chemistry LLM and state-aware chemistry LLM make different decisions at the first route divergence. For the same candidate set, the reference-compatible action is ranked 5th by both the Reviewer and the Base chemistry LLM. The Base Chemistry LLM therefore selects another higher-ranked action, which introduces a structurally implausible disconnection. Although the resulting search reaches purchasable starting materials, the final route does not match the reference synthesis.

In contrast, the state-aware Analyzer promotes the reference-compatible action from rank 5 to rank 1. Following this decision, the subsequent search successfully recovers the reference matched route. This example illustrates how state-aware reaction ranking can correct an early branch selection that would otherwise lead the search toward a different route, highlighting the importance of incorporating the evolving search state into sequential decision making for multi-step retrosynthesis.

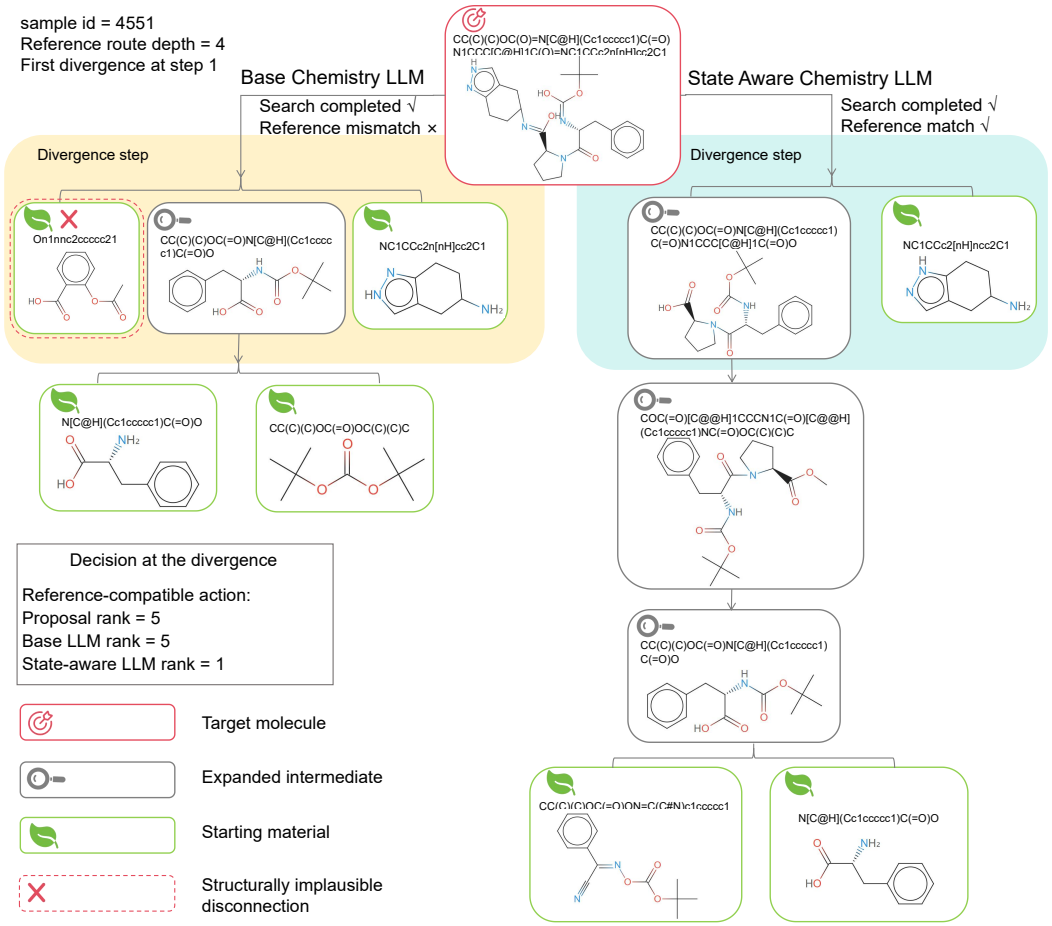


Figure 3: **Case study of route divergence under different reaction rankings.** At the first divergence step, the state-aware Analyzer promotes the reference-compatible action from rank 5 to rank 1, leading to a reference-matched synthesis route.

6 CONCLUSION

We presented RetroRoute, a route-level agentic framework for multistep retrosynthesis that coordinates planning scaffolded reaction proposal, utility guided policy refinement, and search state-aware LLM decision making within an explicit search process. By separating candidate support from state dependent branch evaluation, RetroRoute provides a unified way to connect local reaction generation with route-level planning objectives, addressing the mismatch between single step prediction and multistep search discussed in prior work (Hassen et al., 2022). The framework preserves explicit route construction while allowing proposal generation and branch selection to incorporate complementary planning information. Nevertheless, performance remains constrained on very deep and sparsely represented routes, where both data scarcity and accumulated decision errors become more pronounced. In addition, exact match measures agreement with recorded reference starting materials and may undercount chemically valid alternative syntheses, reflecting a broader limitation of single reference retrosynthesis evaluation (Zagribelnyy et al., 2026). Future work should therefore develop broader route quality evaluation and improve decision learning for longer and less frequently observed planning horizons.

REFERENCES

Andres M Bran, Théo A Neukomm, Daniel P Armstrong, Zlatko Jončev, and Philippe Schwaller. Revealing chemical reasoning in llms through search on complex planning tasks. In *AI for Accelerated Materials Design-ICLR 2025*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. ICML '07, pp. 129–136, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937933.

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

Yujie Chen, AJie Lin, Tengfei Ma, Shu Wu, Leyi Wei, Yiping Liu, and Xiangxiang Zeng. Toward synthesizability-aware multi-step retrosynthetic planning. a.

Yujie Chen, Tengfei Ma, Zhou Yu, Jiayi Zhang, Dongsheng Cao, Yiping Liu, et al. Learning chemical knowledge from large-scale unlabeled molecular data for retrosynthesis. b.

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, 12 2017. ISSN 2374-7943.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Emma Granqvist, Rocío Mercado, and Samuel Genheden. Retrosynformer: planning multi-step chemical synthesis routes via a decision transformer. *Digital Discovery*, 5(1):348–362, 01 2026. ISSN 2635-098X.

Jiasheng Guo, Chenning Yu, Kenan Li, Yijian Zhang, Guoqiang Wang, Shuhua Li, and Hao Dong. Retrosynthesis zero: Self-improving global synthesis planning using reinforcement learning. *Journal of Chemical Theory and Computation*, 20(11):4921–4938, 06 2024. ISSN 1549-9618.

Alan Kai Hassen, Paula Torren-Peraire, Samuel Genheden, Jonas Verhoeven, Mike Preuss, and Igor Tetko. Mind the retrosynthesis gap: Bridging the divide between single-step and multi-step retrosynthesis prediction. In *NeurIPS 2022 Workshop on AI for Science: Progress and Promises*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. Zinc: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52 (7):1757–1768, 07 2012. ISSN 1549-9596.

Chenglong Kang, Xiaoyi Liu, and Fei Guo. Retrointext: A multimodal large language model enhanced framework for retrosynthetic planning via in-context representation learning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 71919–71936, 2025.

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

Hanbum Ko, Chanhui Lee, Ye Rin Kim, Rodrigo Hormazabal, Sehui Han, Sungbin Lim, and Sungwoong Kim. Retroreasoner: A reasoning llm for strategic retrosynthesis prediction, 2026.

Seul Lee, Karsten Kreis, Srimukh Veccham, Meng Liu, Danny Reidenbach, Saee Paliwal, Weili Nie, and Arash Vahdat. Exploring synthesizable chemical space with iterative pathway refinements. In *International conference on learning representations*, volume 2026, pp. 7497–7519, 2026.

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, 10 2017. ISSN 2374-7943.

Gang Liu, Michael Sun, Wojciech Matusik, Meng Jiang, and Jie Chen. Multimodal large language models for inverse molecular design with retrosynthetic planning. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 41744–41771, 2025a.

Gang Liu, Michael Sun, Wojciech Matusik, Meng Jiang, and Jie Chen. Multimodal large language models for inverse molecular design with retrosynthetic planning. In *International Conference on Learning Representations*, volume 2025, pp. 41744–41771, 2025b.

Songtao Liu, Zhengkai Tu, Minkai Xu, Zuobai Zhang, Lu Lin, Rex Ying, Jian Tang, Peilin Zhao, and Dinghao Wu. Fusionretro: molecule representation fusion via in-context learning for retrosynthetic planning. In *International Conference on Machine Learning*, pp. 22028–22041. PMLR, 2023.

Songtao Liu, Hanjun Dai, Yue Zhao, and Peng Liu. Preference optimization for molecule synthesis with conditional residual energy-based models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024a.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, Peter J. Liu, and Xuanhui Wang. LiPO: Listwise preference optimization through learning-to-rank. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2404–2420, Albuquerque, New Mexico, April 2025c. Association for Computational Linguistics. ISBN 979-8-89176-189-6.

Wei Liu, Jiangtao Feng, Hongli Yu, Yuxuan Song, Yuqiang Li, Shufei Zhang, LEI BAI, Wei-Ying Ma, and Hao Zhou. Retro-r1: Llm-based agentic retrosynthesis. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 70709–70737. Curran Associates, Inc., 2025d.

Xiaoyi Liu, Chengwei Ai, Hongpeng Yang, Ruihan Dong, Jijun Tang, Shuangjia Zheng, and Fei Guo. Retrocaptioner: beyond attention in end-to-end retrosynthesis transformer via contrastively captioned learnable graph representation. *Bioinformatics*, 40(9):btae561, 09 2024b. ISSN 1367-4811.

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, 02 2025. ISSN 0022-2623.

Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin Segler. Re-evaluating retrosynthesis algorithms with syntheseus. In *NeurIPS 2023 AI for Science Workshop*, 2023.

Utkarsh Aashu Mishra, David He, Yongxin Chen, and Danfei Xu. Compositional diffusion with guided search for long-horizon planning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Raja S Pappala, Shreyas Vinaya Sathyanarayana, Ronit kumar choudhary, Arjun Verma, and Deepak Warrier. Retrospect: Retrosynthesis via sequential prediction, and chemically transformed-ranking, 2026.

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, 07 2021. ISSN 1549-9596.

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, 09 2019. ISSN 2374-7943.

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.

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.

Yu Shee, Anton Morgunov, Haote Li, and Victor S. Batista. Directmultistep: Direct route generation for multistep retrosynthesis. *Journal of Chemical Information and Modeling*, 65(8):3903–3914, 04 2025. ISSN 1549-9596.

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 (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8818–8827. PMLR, 13–18 Jul 2020.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9405–9415. Curran Associates, Inc., 2021.

Xiaozhuang Song, Xuanhao Pan, Xinjian Zhao, Hangting Ye, Shufei Zhang, Jian Tang, and Tianshu Yu. Aot*: Efficient synthesis planning via llm-empowered and-or tree search. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34727–34758, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1.

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(3):558–572, 03 2024. ISSN 2635-098X.

Austin Tripp, Krzysztof Maziarz, Sarah Lewis, Marwin Segler, and José Miguel Hernández Lobato. Retro-fallback: retrosynthetic planning in an uncertain world. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 30687–30744, 2024.

Zhengkai Tu, Sourabh J. Choure, Mun Hong Fong, Jihye Roh, Itai Levin, Kevin Yu, Joonyoung F. Joung, Nathan Morgan, Shih-Cheng Li, Xiaoqi Sun, Huiqian Lin, Mark Murnin, Jordan P. Liles, Thomas J. Struble, Michael E. Fortunato, Mengjie Liu, William H. Green, Klavs F. Jensen, and Connor W. Coley. Askcos: Open-source, data-driven synthesis planning. *Accounts of Chemical Research*, 58(11):1764–1775, 06 2025. ISSN 0001-4842.

Haorui Wang, Jeff Guo, Lingkai Kong, Rampi Ramprasad, Philippe Schwaller, Yuanqi Du, and Chao Zhang. LLM-augmented chemical synthesis and design decision programs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 62980–63001. PMLR, 13–19 Jul 2025.

Mianchu Wang and Giovanni Montana. Retrosynthesis planning via worst-path policy optimisation in tree-structured mdps. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

YiFei Wang, Qizhi Pei, Jiangtao Feng, Yuntian Shi, Yi Duan, Lihao Wang, Lei Bai, Lijun Wu, Wei-Ying Ma, and Hao Zhou. r^3: End-to-end reasoning-based planning for multi-step retrosynthesis via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37618–37632, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6.

Jiacheng Xiong, Wei Zhang, Yinquan Wang, Jiatao Huang, Yuqi Shi, Mingyan Xu, Manjia Li, Zunyun Fu, Xiangtai Kong, Yitian Wang, Zhaoping Xiong, and Mingyue Zheng. Bridging chemistry and artificial intelligence by a reaction description language. *Nature Machine Intelligence*, 7(5): 782–793, 2025. ISSN 2522-5839.

Nguyen Xuan-Vu, Daniel P Armstrong, Zlatko Jončev, and Philippe Schwaller. Tempre: Template generation for single and direct multi-step retrosynthesis, 2025.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11248–11258. Curran Associates, Inc., 2020.

Kevin Yu, Jihye Roh, Ziang Li, Wenhao Gao, Runzhong Wang, and Connor W. Coley. Double-ended synthesis planning with goal-constrained bidirectional search. In *Advances in Neural Information Processing Systems*, volume 37, 2024. Spotlight.

Yemin Yu, Ying Wei, Kun Kuang, Zhengxing Huang, Huaxiu Yao, and Fei Wu. Grasp: Navigating retrosynthetic planning with goal-driven policy. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10257–10268. Curran Associates, Inc., 2022.

Luotian Yuan, Yemin Yu, Ying Wei, Yongwei Wang, Zhihua Wang, and Fei Wu. Active retrosynthetic planning aware of route quality. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 51666–51685, 2024.

Bogdan Zagribelnyy, Ivan Ilin, Maksim Kuznetsov, Nikita Bondarev, Roman Schutski, Thomas MacDougall, Rim Shayakhmetov, Zulfat Miftahutdinov, Mikolaj Mizera, Vladimir Aladinskiy, Alex Aliper, and Alex Zhavoronkov. When single answer is not enough: Rethinking single-step retrosynthesis benchmarks for llms, 2026.

Situo Zhang, Hanqi Li, Lu Chen, Zihan Zhao, Xuanze Lin, Zichen Zhu, Danyu Luo, Bo Chen, Xin Chen, and Kai Yu. Retrodfm-r: Reasoning-driven retrosynthesis prediction with large language models via reinforcement learning, 2026.

Wei Zhang, Qinggong Wang, Xiangtai Kong, Jiacheng Xiong, Shengkun Ni, Duanhua Cao, Buying Niu, Mingan Chen, Yameng Li, Runze Zhang, Yitian Wang, Lehan Zhang, Xutong Li, Zhaoping Xiong, Qian Shi, Ziming Huang, Zunyun Fu, and Mingyue Zheng. Fine-tuning large language models for chemical text mining††electronic supplementary information (esi) available. *Chemical Science*, 15(27):10600–10611, 2024. ISSN 2041-6520.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Kai Yu, and Xin Chen. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4):102523, 2025. ISSN 2666-3864.

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling*, 60, 12 2019.

Shuangjia Zheng, Tao Zeng, Chengtao Li, Binghong Chen, Connor W Coley, Yuedong Yang, and Ruibo Wu. Deep learning driven biosynthetic pathways navigation for natural products with bionavi-np. *Nature Communications*, 13(1):3342, 2022.

Yanqiao Zhu, Jingru Gan, Xiaoqi Sun, Fang Sun, Yidan Shi, Md Mofijul Islam, Chao Shang, Wenhao Gao, Connor W. Coley, Yizhou Sun, and Wei Wang. Retroagent: Harnessing llms to search over structured memory for agentic retrosynthesis planning, 2026.

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

A IMPLEMENTATION DETAILS

A.1 PLANNING-SCAFFOLDED REACTION PROPOSAL TRAINING

We instantiate the Reviewer with MolT5 and initialize it from the same pretrained checkpoint used in our RetroInText based implementation. To isolate the effect of proposal adaptation and planning context, we construct three proposal variants.

The *Pretrained Proposal* directly uses the initialization checkpoint without additional training on the extracted RetroBench single step corpus.

The *Current Only SFT* model applies Supervised Fine Tuning (SFT) using only the molecule currently selected for decomposition. Its input format is

> *Please predict the reactant of the product:*
> *<current molecule>*

The *Planning Scaffold SFT* model is independently initialized from the same MolT5 checkpoint and trained on the same single step reaction corpus. It replaces the local reaction prompt with the planning scaffold

> *Please predict the reactant of the product:*
> *<target> ...*
> *<current> ...*
> *<current_depth> ...*
> *<max_depth> ...*
> *<goal> purchasable_starting_materials*

The tokens `<target>`, `<current>`, `<current_depth>`, `<max_depth>`, and `<goal>` are added to the tokenizer vocabulary for Planning Scaffold SFT. The maximum depth is fixed to the search budget of 14.

Current Only SFT and Planning Scaffold SFT share identical optimization settings. The detailed training configuration is summarized in Table 3.

A.2 URPO POLICY REFINEMENT

URPO starts from the best Planning Scaffold SFT checkpoint. For each training context, the proposal policy generates up to 20 candidate reactions. Nonreference candidates are evaluated using molecular validity, forward reaction consistency, route compatibility, reference alignment, building block availability, and an undesirable action penalty.

Following the utility formulation in Section 3.3, the implementation uses the weights

$$\lambda_{\text{chem}} = 0.5, \quad \lambda_{\text{fwd}} = 1.0, \quad \lambda_{\text{route}} = 0.5, \tag{12}$$

$$\lambda_{\text{ref}} = 0.3, \quad \lambda_{\text{stock}} = 0.2, \quad \lambda_{\text{bad}} = 1.0. \tag{13}$$

For generated nonreference actions, the reported configuration requires structural validity, zero undesirable action penalty, forward plausibility of at least 0.5, and total utility

$$U(c; x) \geq 1.5. \tag{14}$$

At most 1 utility supported action is retained for each original training example. The total number of generated positive examples is further restricted to at most 0.5 times the number of gold examples. This constraint preserves the demonstrated reaction distribution while allowing supported alternative actions to receive additional policy mass.

The resulting *URPO Refined Proposal* continues optimization from Planning Scaffold SFT. Its training configuration is summarized in Table 3.

Setting	Current Only SFT	Planning Scaffold SFT	URPO Refined Proposal
Initialization	Pretrained MolT5	Pretrained MolT5	Planning Scaffold SFT
Maximum epochs	40	40	30
Learning rate	5×10^{-5}	5×10^{-5}	3×10^{-6}
Batch size	4	4	16
Gradient accumulation	1	1	4
Weight decay	0.1	0.1	0.0
Optimizer	AdamW	AdamW	AdamW
LR schedule	Linear	Linear	Linear
Warmup ratio	0.03	0.03	0.03
Maximum source length	512	512	512
Maximum target length	256	256	256
Maximum gradient norm	1.0	1.0	1.0
Label smoothing	0.0	0.0	0.02
Precision	FP16	FP16	FP16
Early stopping patience	5	5	–
Minimum improvement	10^{-4}	10^{-4}	–
Random seed	42	42	42

Table 3: Training configuration of the Reviewer variants.

A.3 Search-State-Aware Branch Decision Adaptation

We instantiate the Analyzer with ChemDFM. The pretrained model is adapted to state conditioned branch evaluation using Low Rank Adaptation (LoRA).

For each search decision, the Analyzer receives the final target, current molecule, current depth, maximum search depth, remaining search budget, unresolved frontier, recent route history, and up to 10 candidate reactions. Candidate descriptors additionally include the proposal rank, structural validity, and purchasable reactant count. Route history is truncated to the most recent 6 steps.

Candidate actions are represented using symbolic identifiers. Their display order is randomized during training so that the model cannot associate a fixed identifier or position with the supervision label. The LLM outputs the identifier of the selected candidate, and candidate scores are obtained from the corresponding output token logits. Training uses the listwise objective defined in Section 3.4. ChemDFM is adapted with LoRA under 4-bit NF4 quantization and gradient checkpointing.

Only LoRA parameters are optimized. The training configuration is summarized in Table 4. Model selection is based primarily on validation Hit@1, with validation loss used to break ties.

A.4 Closed Loop Search Configuration

Unless stated otherwise, all multistep experiments use the same search configuration, as summarized in Table 5.

The main RetroRoute search uses a global maximum depth of 14. Greedy DFS analysis uses the ground truth route depth as the expansion limit following the corresponding RetroBench protocol.

B Single Step Proposal Results

To examine whether the gains observed in multistep planning are already reflected in reaction proposal quality, we evaluate four proposal variants on the same held out single step reaction set. Table 6 reports cumulative Top-k reaction accuracy.

Supervised adaptation improves reaction proposal accuracy over the pretrained initialization, while the planning scaffold provides a further gain over Current Only SFT under the same reaction corpus. URPO produces the strongest proposal results across all reported k, including a 3.7% improvement in Top-1 over the Planning Scaffold model.

Setting	Value
Maximum candidates	10
Maximum history steps	6
Maximum input length	2048
Training epochs	3
Learning rate	1×10^{-4}
Optimizer	AdamW
LR schedule	Cosine
Training batch size	1
Evaluation batch size	2
Gradient accumulation	32
Effective batch size	32
Weight decay	0.01
Warmup ratio	0.03
Maximum gradient norm	1.0
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
Early stopping patience	2
Random seed	42

Table 4: Training configuration of the state-aware Analyzer.

Parameter	Value
Proposal candidate count	10
Proposal beam size	10
LLM candidate count	10
Retained actions per expansion	5
Route beam size	5
Evaluation Top-k	5
Maximum search depth	14
Maximum search rounds	14
Maximum number of reactants	5
Maximum route history length	6

Table 5: Main multistep search configuration.

The consistent improvement from Top-1 through Top-10 suggests that URPO affects both the highest ranked reaction and the broader candidate pool. This observation complements the multistep results, where URPO improves route coverage before the state-aware decision model is introduced.

C DETAILED ROUTE DEPTH RESULTS

Figure 2 in the main paper focuses on Top-1 behavior for the major route depth groups. Table 7 provides a compact comparison of both Top-1 and Top-5 accuracy across the complete depth range.

The complete RetroRoute configuration provides the strongest Top-1 result across every populated depth group from 2 to 9. The separation from the proposal only variants becomes particularly visible around depths 4 to 6, which supports the use of route state information when decisions accumulate over multiple reaction steps.

URPO also maintains a clear advantage over the Planning Scaffold Proposal across most major depth groups, especially at Top-5. This pattern is consistent with its role in expanding useful reaction support before search state reranking is applied.

The Base ChemDFM control declines rapidly as route depth increases despite using the same URPO proposal model. This behavior further indicates that pretrained chemical knowledge alone is insufficient for state dependent branch evaluation, while listwise adaptation is important for preserving useful routes as the planning horizon grows.

Proposal model	Top-1	Top-3	Top-5	Top-10
Pretrained Proposal	38.3	50.0	54.3	58.0
Current Only SFT	41.9	56.5	60.9	64.9
Planning Scaffold SFT	44.1	58.8	63.3	67.1
URPO Refined Proposal	**47.8**	**62.0**	**66.4**	**70.2**

Table 6: Single step proposal accuracy (%).

	Current Only		Planning Scaffold		URPO		Base ChemDFM		**RetroRoute**	
Depth	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
2	30.5	45.6	34.9	49.2	38.4	54.1	26.4	49.2	**43.8**	**57.8**
3	20.3	32.8	27.0	39.5	28.7	45.0	14.8	29.6	**35.5**	**49.1**
4	14.2	22.0	15.3	26.4	22.0	34.7	5.7	13.5	**31.6**	**41.2**
5	7.8	12.0	12.5	21.3	16.4	29.7	1.7	5.9	**27.5**	**36.8**
6	8.9	11.2	11.2	15.1	21.8	27.4	1.1	3.4	**30.2**	**38.0**
7	0.0	1.4	2.7	5.4	9.5	12.2	0.0	0.0	**12.2**	**17.6**
8	0.0	2.9	0.0	0.0	0.0	0.0	0.0	0.0	**2.9**	**8.8**
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**15.4**	**15.4**
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 7: Top-1 and Top-5 route exact match (%) across route depths.

Results beyond depth 8 involve very small test subsets and should therefore be interpreted cautiously. They are included to provide complete benchmark coverage rather than to support claims about long route behavior.

D GREEDY DFS DIAGNOSTICS

Table 8 reports Greedy DFS accuracy for the controlled RetroRoute variants. The final two rows summarize overall exact match and search success across the complete test set.

RetroRoute achieves the highest Greedy DFS exact match among all controlled variants and preserves its advantage across the main route depth groups. Since Greedy DFS retains only one local action, the result provides additional evidence that the learned state-aware decision model improves branch ordering rather than relying only on beam recovery.

The comparison between search success and exact match is also informative. Base ChemDFM reaches a relatively high search success rate of 88.75%, yet its overall exact match remains only 17.28%. Reaching purchasable terminal molecules therefore does not imply that the selected branches recover a reference-compatible route.

RetroRoute improves both measures, with 90.84% search success and 36.67% exact match. This gap relative to Base ChemDFM supports the claim that listwise search state adaptation improves which complete routes are reached, not merely whether a search can terminate.

E SEARCH BEHAVIOR AND RUNTIME

Figure 4 summarizes the recorded model invocation counts and observed wall clock evaluation time for all controlled configurations.

Depth	Current Only	Planning Scaffold	URPO	Base ChemDFM	**RetroRoute**
2	28.13	33.09	38.92	24.81	**42.45**
3	20.00	25.80	30.40	15.33	**34.93**
4	13.82	15.46	26.13	6.70	**31.46**
5	8.09	12.01	16.67	4.17	**26.23**
6	7.82	8.38	20.11	1.68	**29.61**
7	0.00	2.70	2.70	0.00	**12.16**
8	0.00	0.00	0.00	0.00	**2.94**
9	0.00	0.00	0.00	0.00	**15.38**
10	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00
Overall exact match	21.46	25.92	31.98	17.28	**36.67**
Success rate	81.04	84.43	89.04	88.75	**90.84**

Table 8: Greedy DFS exact match (%) and success rate (%).

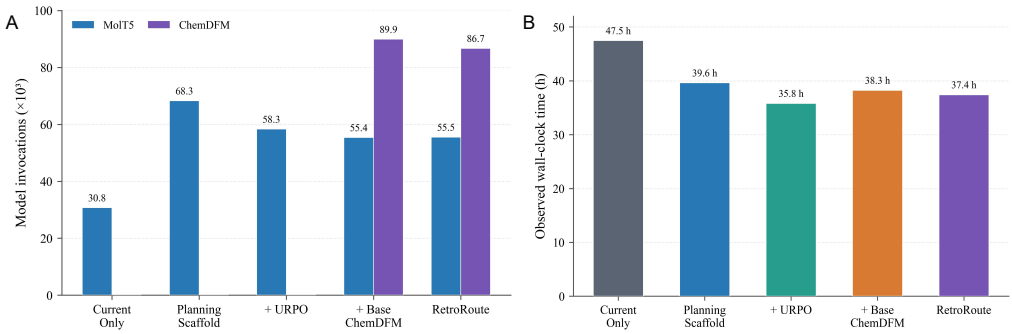


Figure 4: Model invocation counts and observed wall clock evaluation time.

The planning scaffold configuration produces $68,290$ proposal calls, while URPO reduces this number to $58,342$. RetroRoute uses $55,526$ proposal calls, suggesting that changes in proposal quality and branch ranking alter the states visited during search and the point at which routes terminate.

Invocation count is not directly proportional to wall clock time. For example, the Current Only configuration records fewer proposal calls but requires the longest observed evaluation time, while the URPO configuration has the shortest recorded time among the five settings. This difference reflects the interaction between caching, model type, state expansion, and per call computation.

The complete RetroRoute evaluation requires approximately 37.4 hours under the recorded experimental setup while also providing the strongest route accuracy among the controlled variants. We therefore treat Figure 4 as a diagnostic of search behavior and practical evaluation cost rather than as a hardware normalized efficiency comparison.