Retro-Expert: Collaborative Reasoning for Interpretable Retrosynthesis

Anonymous ACL submission

Abstract

Retrosynthesis prediction aims to infer the reactant molecule based on a given product molecule, which is a fundamental task in chemical synthesis. The development of interpretable retrosynthesis models is crucial for chemist's decision by providing meaningful explanation. Building on this, we propose Retro-Expert, an interpretable retrosynthesis reasoning framework that combines domain-specific small models with large language models (LLMs) to generate human-readable reasoning alongside predictions via reinforcement learning. Unlike black-box data-driven models, Retro-Expert outputs natural language explanations grounded in chemical logic (e.g., reaction rules, principles) through three components: (1) specialized small models ensuring chemically valid candidates for reasoning, (2) LLM-driven reasoning to synthesize a decision-making pathway, and (3) reinforcement learning optimizing interpretable decision policy. Experiments show Retro-Expert achieves higher accuracy than single models while producing expert-aligned explanations, bridging AI predictions with actionable chemical insights.

1 Introduction

Retrosynthesis prediction aims to deduce potential reactants and reaction pathways for synthesizing a target product molecule based on its structural characteristics (Somnath et al., 2021; Segler and Waller, 2017; Sun et al., 2021), holding significant application value in drug discovery and molecular design (Hu et al., 2025; Wang et al., 2023, 2018). Existing data-driven models predominantly rely on data memorization mechanisms, which learn mappings between product SMILES and reactant SMILES from datasets, framing the task as either classification or auto-regressive sequence generation (Yao et al., 2023; Chen and Jung, 2021; Yao et al., 2023; Zheng et al., 2019). This paradigm exhibits dual deficiencies: (1) The model can only generate reactant SMILES strings, with no transparency in its internal reasoning process. (2) predictions lack natural language explanations grounded in chemical logic, critically hindering real-world adoption. These limitations lead to a lack of reliable basis for the predictions in chemical principles, which severely undermines chemists' trust in practical applications. Notably, recent breakthroughs (Wang et al., 2025; Xie et al., 2025; Chen et al., 2025) in large language models (LLMs) have demonstrated their potential to address complex specialized problems through specialized-knowledge-based reasoning, enhanced by reinforcement learning via GRPO (Shao et al., 2024). These advancements motivate us to explore how to leverage LLMs' emergent reasoning capabilities to enhance the interpretability of retrosynthesis prediction.

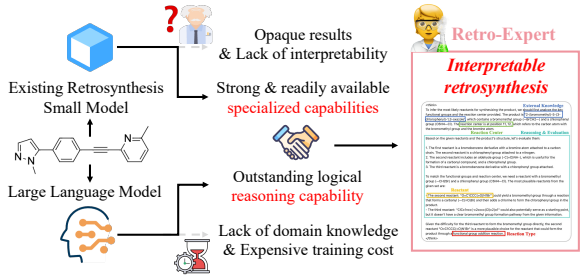

Figure 1: Core highlights of Retro-Expert: the first interpretable retrosynthesis framework capable of generating step-by-step reasoning in natural language, enabled by the collaboration the reasoning capabilities of LLMs with the specialized capabilities of small models.

Therefore, we focus on chemical knowledge-based retrosynthetic reasoning by LLMs to generate reactant results along with explainable reasoning process, ensuring interpretable and transparent retrosynthesis prediction. However, LLMs cannot directly and effectively achieve retrosynthetic reasoning. When employing supervised fine-tuning (SFT) approaches, LLMs typically achieve

retrosynthesis by memorizing common reaction patterns from the training dataset, rather than reasoning based on the underlying chemical principles. In contrast, reinforcement learning (RL) methods can leverage the inherent knowledge of LLMs to simulate expert-like reasoning, enabling the generation of both predictions and interpretable reasoning paths, thus representing a promising direction. However, directly applying RL to incentivize models for retrosynthetic reasoning faces two critical challenges: (1) Domain Knowledge Disparity. Retrosynthesis demands not only logical reasoning but also mastery of specialized chemical knowledge. Pre-trained LLMs fail to adequately internalize and apply specific chemical principles when reasoning solely based on molecular SMILES. (2) Lack of Specialized Capabilities. The retrosynthesis workflow inherently requires coordinated execution of interdependent subtasks, including reaction type classification and reaction center location (Gao et al., 2022; Wang et al., 2021; Yan et al., 2020). Vanilla LLMs struggle to execute these specialized subtasks, while dedicated small models, each individually optimized for a specific subtask, have achieved expert-level performance with no training costs incurred (Wang et al., 2023; Chen and Jung, 2021; Somnath et al., 2021). These small models can offer valuable chemical guidance to LLMs during retrosynthesis. Therefore, we aim to synergistically integrate the specialized capabilities of small models with the advanced reasoning capability of LLMs through reinforcement learning, establishing a new paradigm where small models provide chemical knowledge guidance, upon which LLMs perform explainable decision-making.

Building upon these insights, we present **Exp**lainable and Coop**er**a**t**ive retrosynthesis framework, **Retro-Expert**, the first explainable retrosynthesis framework that integrates natural language-based expert reasoning with model-agnostic compatibility. Retro-Expert achieves dual breakthroughs in prediction accuracy and chemical interpretability by strategically orchestrating small models' domain expertise and the LLM's logical reasoning capabilities. The framework operates through three interconnected modules: (1) **Specialized Multi-Model Candidate Generation.** Leveraging specialized small models (e.g., reaction type classifiers and reactant generators) to produce stage-specific candidate predictions for a target product. (2) **Collaborative Interaction Mechanism between Large and Small Models.** Leveraging their powerful semantic understanding and logical reasoning capabilities, large language models perform integrated analysis and in-depth reasoning over the multi-stage candidate results generated by small models, ultimately producing the final reactant prediction along with interpretable, natural language-based reasoning process. (3) **Knowledge-Constrained Decision Policy Optimization.** Retro-Expert optimizes the large model's reasoning strategy over candidate results via reinforcement learning. A multi-stage reward mechanism is established during training to guide the model toward learning an optimal and trustworthy reasoning path.

Our contributions are summarized as follows:

1. This work represents the first retrosynthesis study capable of generating natural language interpretable reasoning processes. It fills a long-standing interpretability gap in the field, significantly enhancing chemists' trust in the model and its practical applicability in real-world scenarios.
2. We propose Retro-Expert, a collaborative retrosynthesis framework integrating large and small models. It not only improves retrosynthesis prediction accuracy but generates human-understandable step-by-step reasoning processes. Notably, Retro-Expert allows seamless integration of arbitrary small models during inference, enabling flexible expansion without retraining.
3. Systematic experiments validate the advantages of the large-small model collaboration in Retro-Expert. Furthermore, its performance scales with improvements in small model accuracy, demonstrating strong generalization and scalability.

2 Related Work

2.1 Retrosynthesis Prediction

Existing retrosynthesis prediction methods can be broadly categorized into three modeling paradigms: template-based, semi-template-based, and template-free approaches (Gao et al., 2022; Somnath et al., 2021; Sun et al., 2025). Template-based methods (Chen and Jung, 2021; Coley et al., 2017; Dai et al., 2019) apply an appropriate template to the product module via subgraph matching and generate the corresponding reactants. Semi-template-based methods (Yan et al., 2020; Wang et al., 2021; Somnath et al., 2021) first predict the product's reaction center, indicating the location of the leaving group addition for reactants generation. Template-free methods (Sun et al., 2021;

Sacha et al., 2021; Zhong et al., 2022) incorporate auxiliary information such as reaction type (Sacha et al., 2021; Zhang et al., 2024; Liu et al., 2024), to facilitate learning of the direct transformation from product SMILES to reactant SMILES. While each paradigm focuses on distinct aspects of the prediction process to extract intermediate guidance, they often overlook the complementarity and synergy among this intermediate information. Moreover, most existing methods only output the final reactant predictions, which limits the interpretability of the models.

2.2 Large Language Model Reasoning

In recent years, large language models' deep reasoning has been developed to solve specialized scientific problems (Liu et al., 2023; Su et al., 2025; Tang et al., 2025; Putri et al., 2025; Pan et al., 2025). Some studies have shown that a small amount of supervised fine-tuning (SFT) data can enhance a model's reasoning abilities (Zhou et al., 2023; Huang et al., 2025).

However, SFT primarily aims to memorize common patterns from existing datasets to replicate successful reasoning strategies, and its ability to handle more complex tasks remains limited. As a result, reinforcement learning (RL)-based reasoning models like DeepSeek-R1 (Guo et al., 2025; Zhou et al., 2024; Ziegler et al., 2019; Feng et al., 2025) have recently achieved significant progress (Guo et al., 2025; Zhou et al., 2024; Ziegler et al., 2019). Furthermore, some approaches (Team et al., 2025; Chu et al., 2025) have demonstrated the promise of combining long-chain reasoning with reinforcement learning to address highly complex problems, indicating that a strategy that integrates SFT with reinforcement learning may offer a flexible and effective pathway to solve more challenging tasks.

3 Methodology

3.1 Overview

This paper presents Retro-Expert, an interpretable retrosynthesis framework that synergistically combines the complementary strengths of both small and large language models, aiming to enhance the accuracy of retrosynthesis predictions while generating human-understandable reasoning processes. As illustrated in Figure 2, Retro-Expert primarily comprises three core stages. Given the target product, pre-trained specialized small retrosynthesis models are employed to generate candidate predictions for different sub-tasks in retrosynthesis. These candidates serve as the knowledge foundation for the large model's in-depth reasoning (Specialized Multi-Model Candidate Generation). Leveraging the large model's logical reasoning capabilities, cross-stage analysis is performed on the multi sub-task candidate results. Through multi-step reasoning, the optimal results for each sub-task are identified, and a complete natural language reasoning chain is constructed (Collaborative interaction mechanism between large and small models). Using the reinforce learning combined with a multi-stage rule-based reward mechanism and via GRPO (Shao et al., 2024), feedback signals are generated by comparing the large model's intermediate reasoning results with ground-truth labels. This optimizes the large model's reasoning strategy, guiding it to autonomously learn the optimal retrosynthesis reasoning path (Knowledge-Constrained Decision Policy Optimization module).

3.2 Task Definition

The objective of the retrosynthesis task T_{retro} is to predict the set of reactants $\{M_r^i\}_{i=1}^{C}(C \geq 1)$ corresponding to a target product M_p, where $M_p, M_r \in S$, and S represents the valid SMILES space. In real-world expert retrosynthesis prediction, this process typically involves solving n logically connected subtasks $T_{\text{retro}} = \{T_0, T_1, \ldots, T_n\}$, where collaboration among subtasks enables expert-level accuracy. Given an environment $\mathcal{E}$ containing $N(N \geq n)$ specialized models $\mathcal{M} = \{m_0, m_1, \ldots, m_N\}$, each dedicated to a retrosynthesis subtask, the input product M_p is processed by each model m_i to generate Top-K candidate predictions P_i for its corresponding subtask:

$$P_i = \{P_i^k\}_{k=1}^{K}, \quad P_i^k \sim p(m_i|M_p;\theta), \quad (1)$$

where θ denotes the parameters of model m_i, and P_i^k represents the k-th candidate result. The total inference path space $\mathcal{T}$ is defined as the Cartesian product of candidate results from all subtasks:

$$\mathcal{T} = (P_0, P_1, \ldots, P_n), \quad (2)$$

with a space size of K^n. Retrosynthesis reasoning is modeled as a sequential decision-making process in this path space: a LLM M_{LLM} interacts with $\mathcal{E}$, analyzes the Top-K candidates from models in $\mathcal{M}$, and sequentially selects the correct answer $P_i'(P_i' \in P_i)$ for each subtask. This generates a reasoning path $\mathcal{T}_{\text{LLM}} = (P_0', P_1', \ldots, P_n')$ alongside a natural-language explanation R. Retro-Expert has two core

objectives: generating the correct set of reactants $\hat{a} = \{M_r^i\}_{i=1}^C$, and identifying the optimal path T^* with the highest reward within $\mathcal{T}$:

$$\begin{aligned} &\arg\max_{T_{\text{LLM}} \in \mathcal{T}} \text{Reward}(T_{\text{LLM}}) \\ \text{s.t.} \quad &M_{\text{LLM}}(M_p, T_{\text{LLM}}) = \hat{a}. \end{aligned} \tag{3}$$

Here, Reward$(\cdot)$ is a function evaluating the quality of the reasoning path. This framework improves the accuracy of the explainable reasoning R by simultaneously ensuring correct reactant prediction and maximizing the path reward—a critical enhancement not addressed in prior work.

3.3 Specialized Multi-Model Candidate Generation Module

Following the actual retrosynthetic reasoning process of chemists and effectively integrating the domain knowledge required, we propose the Specialized Multi-Model Candidate Generation module. The core concept is to leverage existing pre-trained retrosynthesis models tailored for different subtasks to generate initial candidate results for each sub-task in the retrosynthesis process.

Specifically, the module invokes pre-trained specialized small models (e.g., reaction type prediction models, reaction center localization models) to output candidate results for sub-tasks such as possible retrosynthesis reaction types and reaction center positions of M_p. Since all employed models are pre-trained retrosynthesis models, they can easily generate predictions for each sub-task without additional training costs, significantly enhancing the scalability and usability of Retro-Expert. Notably, the candidate results generated by sub-task models exhibit high recall: their Top-K predictions already contain most correct answers, thus providing sufficient prior knowledge support for the large model's subsequent in-depth reasoning.

Furthermore, the module is highly flexible in design. It not only supports the integration of small retrosynthesis models of any type (e.g., reactant generation models, functional group analysis models) but also incorporates external domain knowledge k (e.g., natural language descriptions of product functional groups, known reaction rules), further enriching the diversity and comprehensiveness of candidate information. Ultimately, the candidate results from all sub-task models and external knowledge are aggregated into a unified information set, which is input to the next stage for collaborative reasoning.

3.4 Collaborative Interaction Mechanism between Large and Small Models

Although LLMs exhibit exceptional capabilities in complex logical reasoning, they lack domain-specific knowledge in retrosynthesis tasks, making it difficult to directly predict reactants based solely on the molecular structure of the product. Thus, the core design goal of Retro-Expert is to use candidate results provided by small models as domain knowledge anchors, guiding LLMs to shift from "***directly predicting reactants***" to "***step-by-step reasoning based on candidate sets***", thereby reducing LLMs' reliance on domain knowledge.

The collaborative mechanism between small and large models is implemented through the following key steps: First, based on the logical chain of chemists' actual retrosynthetic reasoning, we order different sub-task models. Simultaneously, we deduplicate and integrate candidate results from models addressing the same sub-task. This process ultimately forms n logically connected, difficulty-increasing candidate result sets $\mathcal{T}$. These candidate sets, combined with external domain knowledge k, collectively construct the reasoning prompt for the LLM. Second, the constructed prompt is input to the LLM, which is then tasked with performing step-by-step analysis and reasoning on each sub-task. The LLM must derive conclusions for the current step (e.g., localize reaction center) by integrating selection results from previous steps (e.g., confirmed reaction type), ultimately outputting predicted answers for each sub-task and an interpretable reasoning basis in natural language.

In this process, the logical ordering of sub-tasks (from simple to complex, global to local) is critical for ensuring the coherence of the reasoning chain. This collaborative large-small model mechanism integrates the efficient domain knowledge acquisition capabilities of small models with the complex logical reasoning capabilities of LLMs. It not only enhances the accuracy of reactant prediction through expert knowledge constraints but also constructs logically coherent and traceable retrosynthetic reasoning paths, significantly improving the interpretability of model decisions.

3.5 Knowledge-Constrained Decision Policy Optimization Module

Inspired by recent advances in leveraging reinforcement learning (RL) to enhance the reasoning capabilities of large language models (LLMs), we

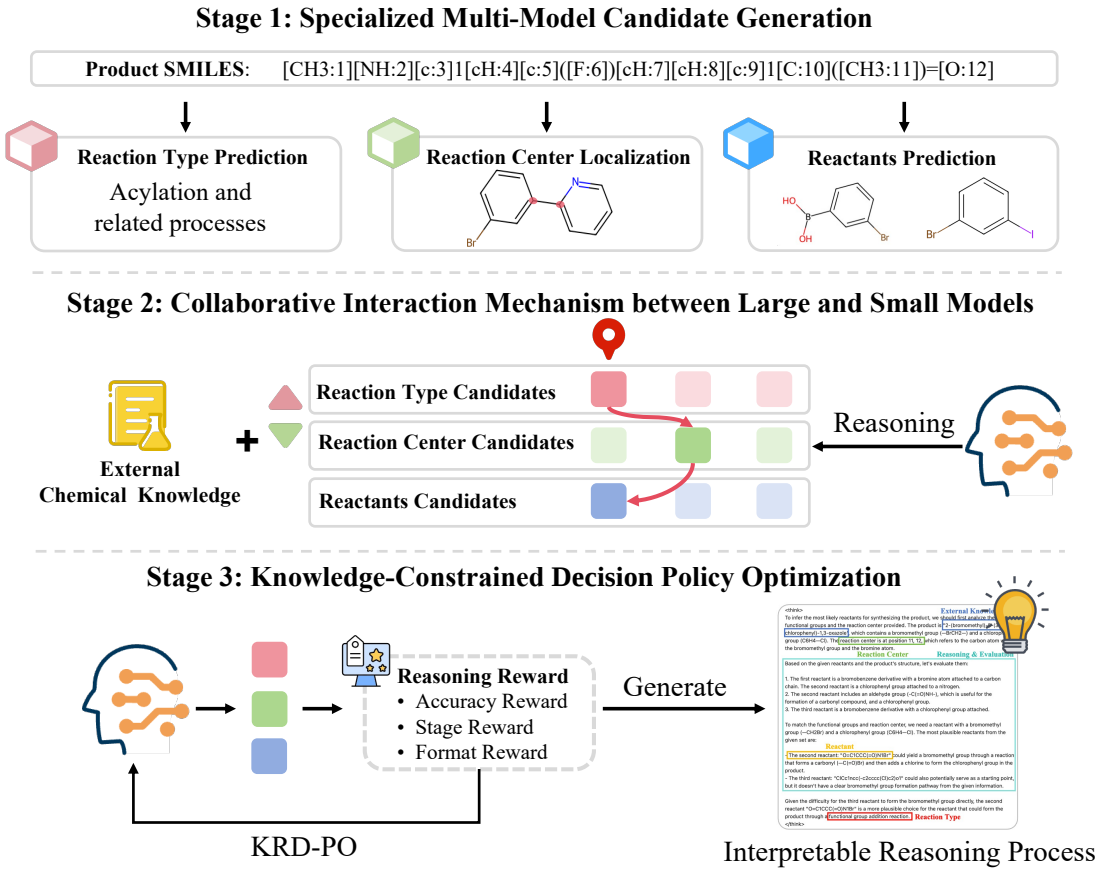

Figure 2: Overview of Retro-Expert. Retro-Expert first invokes multiple specialized small models to perform retrosynthesis analysis on the product and extracts their Top-K predictions as intermediate information to guide the prediction process of the LLM (Stage 1). Based on the candidate results from each model, the LLM then performs multi-step selection and decision-making to adaptively identify a viable retrosynthesis pathway (Stage 2). To optimize the quality and interpretable of the LLM's decision path, we employ KRD-PO to perform end-to-end optimization of the reasoning path.

extend RL to retrosynthetic reasoning to improve the LLM's ability to infer from candidate results. The objective function for this optimization can be formally defined as follows:

$$\max_{\pi_\theta} \mathbb{E}_{q\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|q;\mathcal{E})}\left[r_\phi(q,y)\right] \\ -\beta\, \mathbb{D}_{kl}\left[\pi_\theta(y \mid q;\mathcal{E}) \,\|\, \pi_{\text{ref}}(y \mid q;\mathcal{E})\right] \tag{4}$$

Here, π_θ and π_{ref} denote the policy model and reference model, respectively; r_ϕ represents the reward function, and $\mathbb{D}_{kl}$ is the KL divergence measure. is a question sampled from the dataset (composed of the candidate result set $\mathcal{T}$ and domain knowledge k); y is the model-generated output, encompassing the large model's reasoning path $\mathcal{T}_{LLM}$ and the final reactant prediction result a. To optimize this objective function, we introduce Group Relative Policy Optimization (GRPO) (Shao et al., 2024) and combine it with a rule-based multi-stage reward mechanism to specifically enhance the large model's reasoning capability over the candidate results from retrosynthesis small models.

Compared to previous methods that focus solely on the correctness of the final reactant prediction, we emphasize that the accuracy of the decision-making path during retrosynthetic reasoning is equally critical. In practical applications, chemists' trust in the model stems more from the interpretable reasoning process than from a single result. Thus, relying solely on coarse-grained "answer correctness" rewards is insufficient. We need to design finer-grained reward signals to enhance the authenticity and interpretability of retrosynthesis predictions. Based on this, we propose a Knowledge-Constrained Decision Policy Optimization (KRD-PO) mechanism. KRD-PO seamlessly integrates the domain knowledge of retrosynthesis small models, aiming to maintain overall prediction accuracy while significantly improving the correctness of

retrosynthetic reasoning paths, thereby enhancing the model's interpretability and credibility. Specifically, for a sampled decision path y, the reward function is defined as:

$$r(\mathcal{T}_{\text{LLM}}, y) = \sum_{i=1}^{n} r_i + r_{\text{reactant}} + r_{\text{format}}, \quad (5)$$

where r_i denotes the stage reward for the i-th subtask (where $r_i = 1$ if the prediction matches the sub-task's ground truth label, and 0 otherwise), r_{reactant} is the correctness reward for the final reactant prediction, and r_{format} is the format compliance reward for the reasoning process.

Notably, candidate results generated by small models exhibit a probabilistic bias: the correct answer appears in the Top-1 position with significantly higher probability than in subsequent positions. Without constraints, the LLM might directly select Top-1 results to maximize rewards during training (i.e., the "reward hacking"), leading to a loss of reasoning authenticity. To address this, we pre-shuffle the positions of correct answers in the training samples' candidate results: using a phased decay strategy, we gradually reduce the probability of the correct answer appearing in the top position and increase its distribution in later positions. This operation not only simulates the true Top-K probability distribution of small models but also avoids path selection distortion caused by reward mechanism bias during training. Through KRD-PO optimization, Retro-Expert transcends the limitation of "***accuracy-centric prediction***", instead ***generating coherent, interpretable, and high-quality retrosynthetic reasoning thought chains***. Additionally, it ensures that the reasoning process aligns with chemical principles and expert knowledge, significantly enhancing the model's practical application value in real-world scenarios.

4 Experiments

4.1 Dataset & Evaluation Metric

We conduct experiments using the benchmark dataset USPTO-50K (Schneider et al., 2016) which contains 50,000 atom-mapped reaction records, and split the dataset into training, validation, and test sets with a ratio of 8:1:1 following prior works (Dai et al., 2019; Yan et al., 2020). Based on previous methods (Somnath et al., 2021; Zeng et al., 2024), we standardize the product SMILES and reorder the atom mapping numbers within the product, and reassign atom mapping numbers in the corresponding reactant SMILES. Considering that practical applications typically focus only on the highest-probability prediction, we employ Top-1 accuracy as the metric, defined as the proportion of test cases where the GT reactant appears in the first prediction. Following prior work (Coley et al., 2017; Zheng et al., 2019), we compute the accuracy by comparing the canonical SMILES of predicted reactants to the ground truth.

Category	Model	Top-1 Acc (%)	Natural Language Interpretable
Template-based	LocalRetro	63.9	×
	GLN	64.2	×
Template-free	Retroformer	63.5	×
	UAlign	66.2	×
Semi-Template	GraphRetro	63.9	×
	RetroPrime	64.8	×
	Graph2Edits	67.2	×
Retro-Expert (ours)	+ LocalRetro	**64.1**	✓
	+ UAlign	**67.0**	✓
	+ GraphRetro	**64.1**	✓

Table 1: Top-1 retrosynthesis accuracy (%) on the USPTO-50K test set. Retro-Expert surpasses the performance of small models and achieves natural language interpretability. During inference, Retro-Expert supports ANY small models, allowing them to provide candidate results for sub-tasks without individual training.

4.2 Implementation Details

We employ three specialized small models, including a reaction type prediction model (T5Chem (Lu and Zhang, 2022)), a reaction center prediction model (GraphRetro (Somnath et al., 2021)), and a reactant prediction model (GraphRetro), to provide the necessary information required for expert-level retrosynthesis analysis. During inference, ANY small models can be used to provide candidate results for sub-tasks. To balance accuracy and optimization efficiency, we use Top-3 candidate predictions from each model. We utilize Qwen2.5-3B-Instruct as the LLM and train it with reinforcement learning using only 12k samples. To prevent reward hacking, during training, the correct answers in the small model's candidate predictions are distributed across positions 1, 2, and 3 with a ratio of 5:3:2 to ensure balanced label distribution.

4.3 Comparison Results

For evaluating overall performance, we compare Retro-Expert with existing classic retrosynthesis methods that rely on template-based (LocalRetro (Chen and Jung, 2021), GLN (Dai et al., 2019)), template-free (Retroformer (Yao et al., 2023), UAlign (Zeng et al., 2024))

Strategy	Setting (Train -> Test)	Top-1 Acc	Natural Language Interpretable
SFT	Product -> Reactants	43.2	×
	Product -> Reactants + CoT	32.9	✓
RL	Product -> Reactants	0	✓
	Product + 4 Choice -> Reactants	28.9	✓
	Retro-Expert + GraphRetro	64.1	✓

Table 2: SFT v.s. RL. LLMs lack sufficient chemical knowledge to independently complete retrosynthesis.

Candidate Sets						Top-1 Accuracy			
Type		Center		Reactant		Type	Center	Reactant	Interpretable
R	M	R	M	R	M				
✓		✓		✓		17.4	33.3	29.4	✓
	✓	✓		✓		75.6	36.5	31.0	✓
	✓		✓	✓		**75.6**	**84.5**	**64.1**	✓
	✓		✓		✓	75.6	84.5	63.9	×

Table 3: Performance of different sub-tasks on Reasoning and Memorization. "R" and "M" denotes using reasoning (LLM) or memorization (small model) to provide candidates.

and semi-template-based (GraphRetro (Somnath et al., 2021), RetroPrime (Wang et al., 2021), Graph2Edits (Zhong et al., 2023)).

As shown in Table 1, we selected three representative models of different types to provide candidate results for Retro-Expert. Retro-Expert consistently surpasses the standalone Top-1 accuracy of specialized retrosynthesis models while maintaining proportional performance gains as baseline model accuracy increases. Crucially, our framework uniquely generates chemically grounded natural language explanations, a distinguishing feature absent in conventional methods.

4.4 Ablation Study

Reasoning capabilities of LLMs in retrosynthesis. To validate the necessity of large-small model collaboration, i.e., LLMs require specialized models to provide domain knowledge they lack, we conducted a detailed analysis of LLMs' reasoning capabilities in retrosynthesis.

As shown in Table 2, we trained LLMs using Supervised Fine-Tuning (SFT) and RL respectively. When predicting reactants with only the product as input, the SFT model obtains answers by memorizing training data rather than performing reasoning. We further used Deepseek-V3 to generate Chain-of-Thought (CoT) reasoning processes from products to reactants for the training set, and finetuned the model. Although the LLM could now output reasoning processes, its Top-1 accuracy decreased by 10.3%, primarily due to numerous chemical factual errors in the CoT obtained by Deepseek-V3. To further verify the model's internal chemical knowledge, we trained the model using RL with a rule-based reward mechanism to directly predict reactants from the product. However, all test predictions were incorrect, indicating that the LLM lacks sufficient knowledge to independently solve retrosynthesis tasks. When provided with the product and 4 candidate reactants, the model's Top-1 accuracy increased to 28.9%, yet this remains significantly lower than Retro-Expert (46.1%). This demonstrates that LLMs require specialized models to provide domain knowledge for each stage of retrosynthetic reasoning.

Top-K Candidates	1	2	3	4	5
Reactant Top-1	66.8	65.0	64.1	62.5	60.5
Δ	**+2.9**	**+1.1**	**+0.2**	-1.4	-3.4

Table 4: The effect of the K value in small models' Top-K candidates on large model performance. Δ denotes the improvement relative to the Top-1 accuracy of small models' reactant prediction (63.9%).

Reasoning and memorization in different sub-tasks of retrosynthesis. To validate the respective strengths of small models and large models in retrosynthesis, we compared the performance of large model reasoning versus small model memorization across different sub-tasks. Here, "reasoning" refers to small models providing Top-3 candidates for LLM reasoning, while "memorization" refers to small models directly supplying Top-1 candidates to the LLM. When all sub-tasks relied on LLM reasoning, overall performance was poor, indicating that LLMs struggle to obtain accurate chemical knowledge support. When the reaction type sub-task used small model memorization, reaction type performance improved by 58.2%, with concurrent improvements in reaction center and reactant predictions. This suggests that reaction type classification is better suited for small models to memorize based on product SMILES. Further applying memorization to the reaction center sub-task increased reactant prediction accuracy from 31.0% to 64.1%, indicating that reaction center localization directly impacts reactant prediction performance and provides the most direct information to aid LLM reasoning.

Effectivenss of small models' Top-K candidates. We analyze the impact of using different K values for candidate results during inference on LLM reasoning performance in Table 4. During training, small models only provide 3 candidate results for LLM. As K decreases, the number of

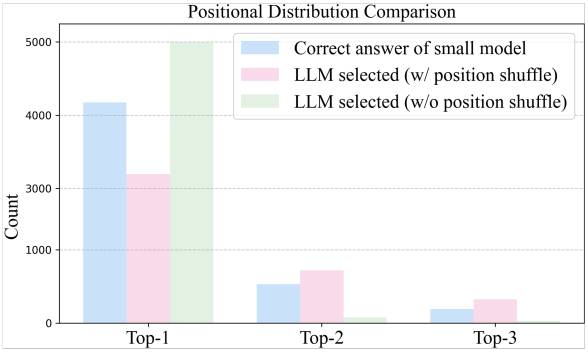

Figure 3: Comparison of positional distributions between the correct answers of small models and the selected positions of LLMs.

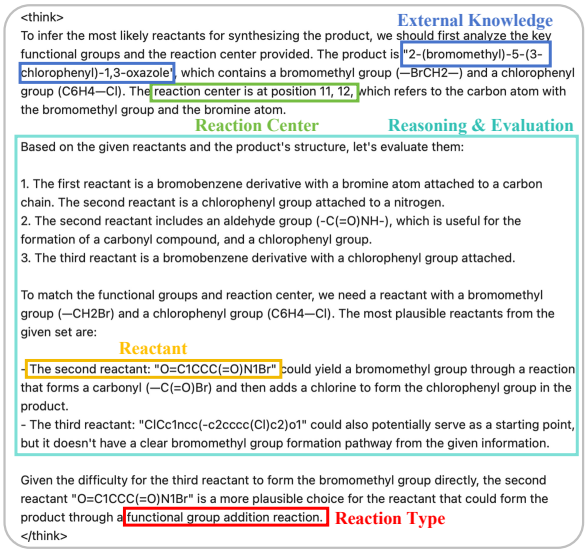

Figure 4: Demonstration of Interpretable Reasoning Process. Retro-Expert guides the LLM to derive reactants via step-by-step reasoning based on candidate results from diverse small models. Such interpretable reasoning processes significantly enhance experts' trust in the model and boost its practical applicability.

candidates the LLM needs to reason over reduces, lowering the reasoning difficulty (the expected correct reasoning probability changes from 33.3% to 50%). Consequently, reactants prediction accuracy gradually increases. We surprisingly found that when K equals to 1, reactant accuracy reached 66.8%, exceeding the small model's Top-1 accuracy by 2.9%. This indicates that even when small models fail to provide the correct answer, the LLM retains the ability to correct its predictions from erroneous candidates. As K increases, the LLM's reasoning difficulty rises, ultimately leading to a 3.4% decrease in accuracy. We attribute this decline to the training setting, increasing the training K value could further force the model to enhance its reasoning capabilities.

Reward hacking. Due to the probabilistic bias in the Top-K candidate results provided by small models, i.e., the correct answer appears in the Top-1 position with significantly higher probability than in other positions. When using result correctness as the reward, the LLM will directly select the first candidate to continuously maximize its reward. To validate the effectiveness of our candidate position shuffling strategy, we visualized the positional distribution of correct answers in the small model's K candidates and the distribution of positions selected by the LLM. As shown in Figure 3, using the original K candidates from small models results in severe reward hacking, whereas our position shuffling strategy effectively mitigates this issue.

4.5 Case Study

In Figure 4, we visualize the natural language-based interpretable reasoning process generated by Retro-Expert. Observations show that small models effectively provided knowledge anchors for the LLM through reaction type and reaction center predictions, prompting the LLM to further leverage external knowledge for reasoning based on domain knowledge. Additionally, the LLM analyzed each candidate's result provided by small models step-by-step and ensured the rationality of the generated results through self-verification.

5 Conclusion

In this paper, we present Retro-Expert, the first interpretable retrosynthesis framework capable of generating step-by-step reasoning in natural language, addressing long-standing interpretability limitations in existing methods. By leveraging reinforcement learning to collaborate small models' specialized chemical expertise with LLMs' advanced reasoning capabilities, Retro-Expert delivers both accurate reactant predictions and step-by-step explanations grounded in chemical logic. Our experiments show that Retro-Expert not only outperforms prior approaches with minimal training data but also accommodates the seamless integration of arbitrary small models during inference. This framework enhances model trustworthiness and bridges the gap between opaque model predictions and chemists' logic-driven workflows, providing a practical tool for retrosynthesis planning.

Limitations

Limited by our computational resources, Retro-Expert was only trained using reinforcement learning on Qwen-2.5-3B. However, its reasoning capabilities can be further enhanced with increasing large language model (LLM) parameter sizes (e.g., 7B or 32B). Although manual observation indicates that the interpretable reasoning chains generated by Retro-Expert are highly accurate, quantitative evaluation remains necessary. However, how to conduct such evaluation poses a significant challenge, as it involves fundamental chemical knowledge, expert reasoning processes, and chemical reaction feasibility. We will continue to explore this direction in future work.

References

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

Yongrui Chen, Junhao He, Linbo Fu, Shenyu Zhang, Rihui Jin, Xinbang Dai, Jiaqi Li, Dehai Min, Nan Hu, Yuxin Zhang, and 1 others. 2025. Pandora: A code-driven large language model agent for unified reasoning across diverse structured knowledge. *arXiv preprint arXiv:2504.12734*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. 2017. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245.

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

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. 2022. Semiretro: Semi-template framework boosts deep retrosynthesis prediction. *arXiv preprint arXiv:2202.08205*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Zhaolin Hu, Yixiao Zhou, Zhongan Wang, Xin Li, Weimin Yang, Hehe Fan, and Yi Yang. 2025. Osda agent: Leveraging large language models for de novo design of organic structure directing agents. In *The Thirteenth International Conference on Learning Representations*.

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

Pengfei Liu, Jun Tao, and Zhixiang Ren. 2024. A self-feedback knowledge elicitation approach for chemical reaction predictions. *arXiv preprint arXiv:2404.09606*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. 2023. Multimodal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*, 62(6):1376–1387.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Rafa Anugrah Putri, Ahmad Taufiq, and 1 others. 2025. Effectiveness of innovative learning models to improve scientific reasoning on physics topics: A literature review. *Jurnal Penelitian Pendidikan IPA*, 11(3):19–22.

Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. 2021. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346.

Marwin HS Segler and Mark P Waller. 2017. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. 2021. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34:9405–9415.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, and 1 others. 2025. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*.

Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. 2021. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34:10186–10194.

Shengyin Sun, Wenhao Yu, Yuxiang Ren, Weitao Du, Liwei Liu, Xuecang Zhang, Ying Hu, and Chen Ma. 2025. Gdiffretro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. *arXiv preprint arXiv:2501.08001*.

Xiangru Tang, Tianyu Hu, Muyang Ye, Yanjun Shao, Xunjian Yin, Siru Ouyang, Wangchunshu Zhou, Pan Lu, Zhuosheng Zhang, Yilun Zhao, and 1 others. 2025. Chemagent: Self-updating library in large language models improves chemical reasoning. *arXiv preprint arXiv:2501.06590*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Jingxue Wang, Huali Cao, John ZH Zhang, and Yifei Qi. 2018. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):1–9.

Junxiong Wang, Wen-Ding Li, Daniele Paliotta, Daniel Ritter, Alexander M Rush, and Tri Dao. 2025. M1: Towards scalable test-time compute with mamba reasoning models. *arXiv preprint arXiv:2504.10449*.

Xiaorui Wang, Yuquan Li, Jiezhong Qiu, Guangyong Chen, Huanxiang Liu, Benben Liao, Chang-Yu Hsieh, and Xiaojun Yao. 2021. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845.

Yiming Wang, Yuxuan Song, Minkai Xu, Rui Wang, Hao Zhou, and Weiying Ma. 2023. Retrodiff: Retrosynthesis as multi-stage distribution interpolation. *arXiv preprint arXiv:2311.14077*.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

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

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, and 1 others. 2023. Retroformer: Retrospective large language agents with policy gradient optimization. *arXiv preprint arXiv:2308.02151*.

Kaipeng Zeng, Bo Yang, Xin Zhao, Yu Zhang, Fan Nie, Xiaokang Yang, Yaohui Jin, and Yanyan Xu. 2024. Ualign: pushing the limit of template-free retrosynthesis prediction with unsupervised smiles alignment. *Journal of Cheminformatics*, 16(1):80.

Xu Zhang, Yiming Mo, Wenguan Wang, and Yi Yang. 2024. Retrosynthesis prediction enhanced by in-silico reaction data augmentation. *arXiv preprint arXiv:2402.00086*.

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. 2019. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling*, 60(1):47–55.

Weihe Zhong, Ziduo Yang, and Calvin Yu-Chian Chen. 2023. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1):3009.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.