Predict-then-Verify: Aligned Graph-Language Models for Multi-Step Chemical Mechanism Prediction

Anonymous ACL submission

Abstract

Anticipating multi-step reaction mechanisms remains a formidable challenge in organic chemistry. Existing computational methods are inherently limited to single-step elementary reaction prediction and struggle to extrapolate to unseen reactions. To bridge this gap, we propose a novel *predict-then-verify* framework for robust, multi-step mechanism pathway reasoning. Our framework couples a Graph-aligned Reaction Mechanism Predictor (GRMP) with a dynamic Relation Retrieval (RR) verifier, underpinned by the newly pioneered Chemical Mechanism Knowledge Graph (CMKG). GRMP is optimized via a two-stage strategy: supervised fine-tuning (SFT) on the shortest CMKG pathways, followed by graph-grounded preference learning to mitigate structural hallucinations. During inference, the RR module leverages each predicted intermediate as a topological anchor to dynamically query the CMKG, extracting factual relation triplets as structural priors to guide subsequent reasoning. Extensive evaluations on the CMQA and oMe-Gold benchmarks demonstrate that our framework achieves state-of-the-art performance. Crucially, under challenging out-of-distribution (OOD) settings, our method achieves a 36.3% pathway accuracy on unseen reactions, showcasing exceptional generalization and scientific plausibility. Code and data are available at `https://anonymous.4open.science/r/CMKG-release/`.

1 Introduction

In organic chemistry, a reaction mechanism is a theoretical trajectory mapping electron redistribution during a chemical transformation. Chemists rely on an understanding of reaction mechanisms as a guiding framework for predicting reaction outcomes, often represented as "arrow pushing" diagrams (Levy, 2017). By sequentially applying arrow-pushing diagrams to the reactants, intermediates are generated until the final products are formed. However, due to the exponentially increasing number of potential pathways, proposing accurate multi-step mechanism pathways remains a challenge even for experienced domain experts.

Traditional methods for mechanism prediction involve three stages: (1) developing rule-based template systems to encode expert knowledge and label mechanistic data (Burés and Larrosa, 2023; Joung et al., 2024; Chen et al., 2024b); (2) retraining end-to-end machine learning models on mechanism data for single-step elementary reaction prediction (Tavakoli et al., 2023a); (3) developing deep learning-based or generative methods to model electron redistribution to adhere to fundamental physical laws for single-step elementary reaction prediction (Joung et al., 2025; Chen et al., 2025). However, these methods are inherently constrained by their reliance on single-step elementary reaction data. Consequently, they are unable to *extrapolate across complex, multi-step mechanistic pathways*, even failing to *generalize to unseen reactions* where the mechanistic logic deviates from known training patterns.

Recent advances in large language models (LLMs) have shown exceptional capabilities in complex reasoning across various chemical domains (Zhang et al., 2025; Li et al., 2026; Gottweis et al., 2026). Despite this progress, complex chemical reasoning, particularly multi-step mechanistic inference, remains largely unexplored, with existing benchmarks predominantly confined to simplified multiple-choice mechanism selection (Hao et al., 2026). This gap stems from a critical bottleneck: LLMs severely struggle with faithful scientific reasoning due to *lack of chemical knowledge* and structural *hallucination* (Qiao et al., 2023; Huang and Chang, 2023; Huang et al., 2024; Wang et al., 2023).

To address these issues, many studies have integrated Knowledge Graphs (KGs), which encapsulate factual information in a structured format,

to improve the reasoning abilities of LLMs. However, existing KG-augmented approaches fail when directly deployed in complex chemical spaces. Extracting comprehensive, multi-hop mechanistic pathways from a chemical graph introduces severe *topological noise*. Conversely, shallow retrieval is insufficient for governing chemical steps, leaving downstream LLMs under-informed and still prone to cascading hallucinations.

To bridge this gap, we construct the Chemical Mechanism Knowledge Graph (CMKG) by imputing intermediates into raw reaction sequences using a set of expert-validated elementary reaction templates. Next, we propose a **predict-then-verify** framework for multi-step mechanism pathway reasoning. As illustrated in Figure 1, our framework couples a Graph-aligned Reaction Mechanism Predictor (GRMP) with a dynamic Relation Retrieval (RR) verifier. To enable robust generation of mechanism pathways, we train GRMP via a two-stage strategy: supervised fine-tuning on the shortest CMKG pathways followed by preference optimization that jointly leverages graph-based link prediction and SFT-generated pathways. During inference, the RR module leverages each predicted intermediate as a topological anchor to dynamically query the CMKG, extracting factual relation triplets as structural priors to guide subsequent reasoning. The candidate pathway and verified triplets are then synthesized by a general-purpose LLM to output final mechanism pathways. Extensive evaluations on the oMe-Gold benchmark and newly curated datasets demonstrate that our framework achieves state-of-the-art predictive performance and exhibits powerful out-of-distribution generalization to unseen reaction classes.

The main contributions of this work are as follows:

- We construct CMKG, a knowledge graph that enriches raw reaction data with imputed intermediates using expert-curated templates.
- We propose a novel predict-then-verify framework, integrating a graph-aligned LLM (GRMP) with an RR module that acts as a structural verifier for mechanism prediction.
- We conduct extensive experiments and human evaluation, demonstrating that GRMP achieves state-of-the-art performance and generalizes effectively to unseen reactions.

2 Related Work

Mechanism prediction. Anticipating reaction mechanism inference at the elementary step level is crucial for predicting complex chemical outcomes. Early work focused on data curation by expert-annotated templates, such as MechFinder (Chen et al., 2024b), which subsequent works expanded to cover wider reaction distributions (Chen et al., 2025; Joung et al., 2024). Driven by these datasets, contemporary research applies deep learning to predict intermediate cascades. Notably, Graph-based models autoregressively generate consecutive intermediates over reaction graphs (Joung et al., 2024). FlowER (Joung et al., 2025) models electron redistribution as a generative process for single-step predictions. Despite their success, these methods fail to autonomously extrapolate across multi-step mechanistic pathways and exhibit poor out-of-distribution (OOD) generalization when facing novel, unseen transformations.

KG-enhanced LLM reasoning. To mitigate the knowledge gap and hallucinations inherent in LLM, incorporating KGs has emerged as a prominent paradigm for reasoning fidelity (Pan et al., 2024). For instance, KD-CoT (Wang et al., 2023) retrieves facts from an external knowledge graph to guide the CoT performed by LLMs. RoG (Luo et al., 2024) proposes a planning-retrieval-reasoning framework that retrieves reasoning paths from KGs to guide LLMs in conducting faithful reasoning. GCR (Luo et al., 2025a) dynamically integrates topological constraints directly into the LLM decoding process to guarantee faithful generation. Besides, GNN-RAG (Mavromatis and Karypis, 2024) and GFM-RAG (Luo et al., 2025b) adopt the graph neural network to learn the relational embeddings for retrieval in KGs. G-reasoner (Luo et al., 2025c) further integrates graph foundation models for scalable reasoning over diverse graph-structured knowledge.

3 Preliminary

Chemical mechanism pathway. A chemical reaction transforms a set of reactants and conditions into final products $\mathcal{P}$. Formally, a mechanism pathway w is defined as a sequence of t elementary steps: $w = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \dots \xrightarrow{r_k} e_k$ represents intermediate molecules transited by the elementary templates written by experts. e_i, r_i represent the molecular entities and relations.

Chemical Mechanism Question Answering

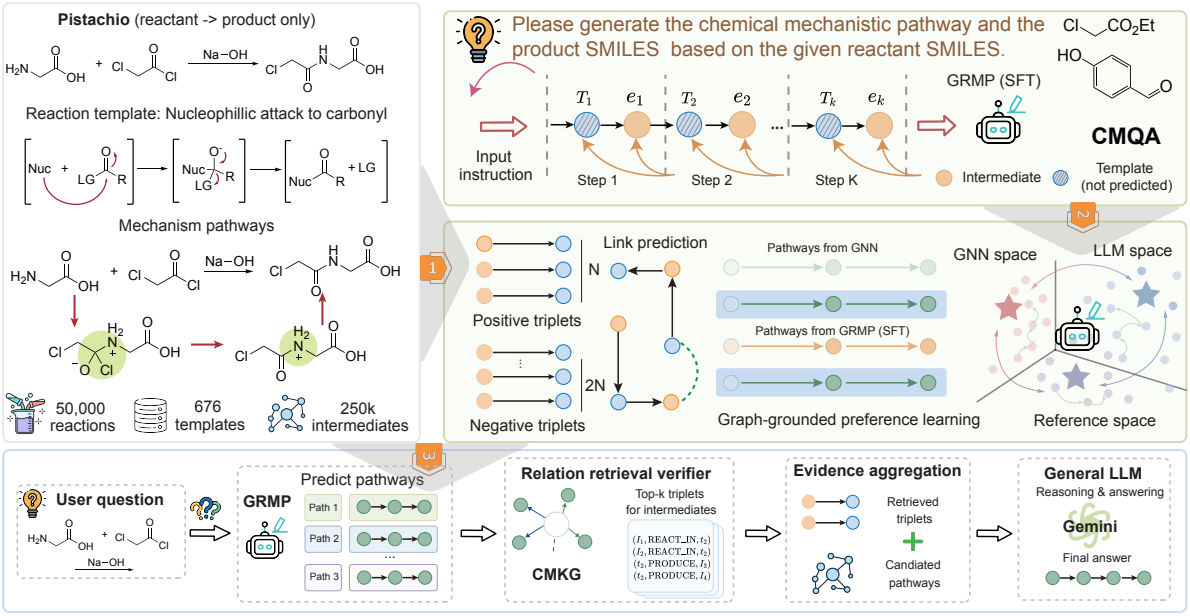

Figure 1: An illustration of our proposed predict-then-verify framework. It contains GRMP, a graph-aligned LLM predictor and a Relation Retrieval (RR) verifier for the mechanism pathway prediction and reasoning.

(CMQA). We frame multi-step mechanism prediction as a specialized QA task over $\mathcal{G}$. Given a natural language query q, equipped with e_0, the objective is to learn a function f that predicts a sequence of sound mechanistic pathways $w \in \mathcal{A}$ anchored in $\mathcal{G}$, such that $w = f(q, \mathcal{G})$.

4 Methods

4.1 Dataset Curation

We use the Pistachio database processed in (NextMove Software), which contains 15 million reactions as the foundation for our mechanistic dataset. Building on the previous methodology for imputing mechanistic pathways (Joung et al., 2024), we derive structured mechanistic pathways from 50,000 experimentally validated reactions using 676 expert-curated reaction templates proposed by (Joung et al., 2025). All mechanistic data are manually and automatically reviewed and cleaned to enhance their consistency and reliability.

After that, we define reaction templates and all molecules as entities, and mechanism subtypes as relations to construct CMKG (Figure 4). This results in 247,832 entities, of which 247,156 are molecule nodes, and 676 are template nodes. The details of the nodes and entities can be found in Table 6 and Table 8. We next extract the shortest reasoning paths, defined as sequences of consecutive triplets: $w_z = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \dots \xrightarrow{r_k} e_k$, where $\forall (e_{i-1}, r_i, e_i) \in \mathcal{G}$ to construct CMQA for model development (Figure 2). We partition these reactions into 40,000 for the training set, 5,000 for validation, and 5,000 for testing. The detailed construction for CMKG is in Appendix A.

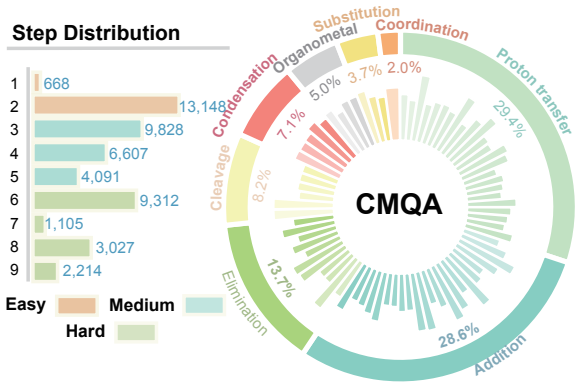

Figure 2: Taxonomy of CMQA. This benchmark spans 13 major categories of organic reactions. We visualize the distribution of the top-8 categories. This diversity ensures a robust evaluation across different domains.

4.2 GNN for Pathway Prediction

To capture the topological dependencies between the input reactant query and the reaction intermediates within the Chemical Mechanism Knowledge Graph (CMKG), we employ a query-dependent Graph Neural Network (GNN) backbone. Based on the definitions in the preliminary, we map each entity node $e \in \mathcal{E}$ to an initial textual embedding $\mathbf{z}_e \in \mathbb{R}^d$ using a pre-trained text embedding model (Chen et al., 2024a).

To inject the specific query context into the graph topology, we designate the reactant nodes present in the input query q as query-relevant nodes $\mathcal{E}_q \subseteq \mathcal{E}$. The query-conditioned initial node representation $\mathbf{h}_e^0$ is defined as:

$$\mathbf{h}_e^0 = \texttt{Init}\left(\mathbf{z}_e,\ \mathbb{1}[e \in \mathcal{E}_q] \cdot \mathbf{z}_q\right), \quad \forall e \in \mathcal{E}. \quad (1)$$

where $\mathbf{z}_q \in \mathbb{R}^d$ is the query embedding derived from the input reactant SMILES, $\mathbb{1}[\cdot]$ denotes the indicator function, and `Init` is a single-layer Multi-Layer Perceptron (MLP) that fuses the intrinsic node embedding with the query vector for query-relevant nodes, while leaving non-query nodes unchanged.

We then perform L layers of message passing over the entity graph to align topology with mechanistic reasoning patterns. At each layer $l \in \{1, \ldots, L\}$, the embedding of node e is updated by aggregating features from its localized neighborhood:

$$\mathbf{h}_e^l = \texttt{Update}\Big(\mathbf{h}_e^{l-1},\ \texttt{Agg}\big(\mathcal{M}_e^l\big)\Big), \quad (2)$$

where $\mathcal{M}_e^l = \left\{\texttt{Msg}\left(\mathbf{h}_e^{l-1}, \mathbf{h}_{e'}^{l-1}\right) \mid (e', e) \in \mathcal{E}\right\}$ represents the set of incoming messages from neighboring entities. The message function `Msg` parameterizes the interaction between connected entities as an element-wise product:

$$\texttt{Msg}(\mathbf{h}_e, \mathbf{h}_{e'}) = \mathbf{h}_e \odot \mathbf{h}_{e'}, \quad (3)$$

where $\odot$ denotes element-wise multiplication. The `Agg` function aggregates these localized neighbor messages, and `Update` non-linearly combines the aggregated neighborhood context with the node's historical representation.

To delineate the entire multi-step trajectory, we frame mechanism prediction as an iterative link prediction task over entity sequences. Starting from the initial reactant node e_0, at each execution step t, the model autoregressively predicts the subsequent intermediate node in the pathway utilizing the final-layer graph embeddings $\mathbf{h}_e^L$:

$$p(e_{t+1} \mid e_t, \mathcal{Q}) = \texttt{Predictor}\left(\mathbf{h}_{e_t}^L, \mathbf{z}_q\right), \quad (4)$$

where `Predictor` is a two-layer MLP topped with a softmax layer over the entire entity space $\mathcal{E}$. The framework greedily selects the next optimal entity via $e_{t+1} = \arg\max_{e \in \mathcal{E}} p(e)$ and appends it to the trajectory. This iteration terminates when either the final product node is reached or a predefined maximum step threshold $T_{\max}$ is exceeded. The predicted pathways are denoted as w^{gnn}.

4.3 Graph-align LLM for Mechanism Prediction (GRMP)

Given a chemical reasoning prompt q and a target reactant entity e_0 identified within q, GRMP's objective is to map the text input to a valid mechanistic pathway w. We utilize the shortest pathway in the CMKG as ground-truth target pathways w^*.

Supervised Fine-Tuning Stage. We first apply SFT (Ouyang et al., 2022) to train the LLM to generate entity paths conditioned on the joint context of prompt q and target entity e_0. This stage forces the LLM to align raw language semantics with discrete entity transitions over $\mathcal{G}$. Let $w^* = (e_0^*, e_1^*, \ldots, e_K^*)$ denote a gold pathway. The SFT objective maximizes the conditional log-likelihood $\mathcal{L}_{\text{SFT}}$ of the target entity sequence:

$$\mathcal{L}_{\text{SFT}}(\theta_s) = -\sum_{t=1}^{K} \log P_{\theta_s}\left(e_t^* \mid e_{<t}^*, q, e_0\right), \quad (5)$$

where θ_s denotes the parameters of the LLM, and $e_{<t}^*$ represents the generated entities.

Graph-Grounded Preference Optimization. To further suppress intermediate-level structural hallucinations and rectify invalid chemical transitions, we introduce a preference optimization stage. We curate a preference dataset $\mathcal{D}_{\text{pref}} = \{(q, w_w, w_l)\}$ by benchmarking the trajectories generated by the GNN predictor (w^{gnn}) and the baseline SFT model (w^{sft}) against the ground-truth path w^*.

We devise a dual-threshold filtering strategy governed by relative improvement α and absolute fidelity β to identify high-quality winning paths (w_w) and losing paths (w_l). Let $\text{sim}(w_a, w_b)$ quantify the similarity between two trajectories evaluated via Tanimoto similarity over their constituent components. The routing metrics are formulated as:

$$\begin{cases} \alpha = \text{sim}(w^{\text{gnn}}, w^*) - \text{sim}(w^{\text{sft}}, w^*) \\ \beta = \text{sim}(w^{\text{sft}}, w^*) \end{cases} \quad (6)$$

Using the empirical threshold value of 0.3, pairs are categorized as follows: *GNN-preferred pairs:* If $\alpha > 0.3$, the GNN prediction exhibits superior topological fidelity; hence, we assign $w_w = w^{\text{gnn}}$ and $w_l = w^{\text{sft}}$. *Gold-preferred pairs:* If $\beta < 0.3$, the SFT prediction deviates severely from chemical reality; we thus assign the gold path as the winner $w_w = w^*$ and $w_l = w^{\text{sft}}$.

This data curation paradigm explicitly forces the model to learn from the failure modes of the SFT

baseline. Let π_θ denote the target policy LLM and π_{ref} represent the frozen SFT reference model. The DPO objective minimizes the following negative log-likelihood loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\,\mathbb{E}_{(q,w_w,w_l)\sim\mathcal{D}_{\text{pref}}}\Bigg[\log\sigma\cdot \Bigg(\lambda\log\frac{\pi_\theta(w_w \mid q, e_0)}{\pi_{\text{ref}}(w_w \mid q, e_0)} - \lambda\log\frac{\pi_\theta(w_l \mid q, e_0)}{\pi_{\text{ref}}(w_l \mid q, e_0)}\Bigg)\Bigg], \tag{7}$$

where σ is the sigmoid function and λ is a temperature hyperparameter scaling the preference constraint. By minimizing $\mathcal{L}_{\text{DPO}}$, GRMP learns to steer its generation away from hallucinatory structural shortcuts and toward chemically plausible reaction mechanisms.

4.4 Relation Retrieval for Inductive Reasoning

During inference, we design a Relation Retrieval module to extract factual triplets serving as structural priors for inductive reasoning. Given an input reaction query q, GRMP generates a candidate mechanism pathway $w = (e_0, e_1, \ldots, e_K)$. For each $e_k \in w$, the module first performs entity linking to map the SMILES-based intermediate to its corresponding node $e_t \in \mathcal{E}$ using the Tanimoto function ($\mathbf{Sim}$).

$$e_t = \arg\max_{e\in\mathcal{E}} \mathbf{Sim}(I_t, e) \tag{8}$$

Upon identifying the structural anchors $E_{\text{anchor}} = \{e_1, e_2, \ldots, e_T\}$, we retrieve its k-hops relation triplets for each anchor node e_t: $\mathcal{N}_k(e_t)$. Finally, we prompt retrieved triplets for the general-purpose LLM to generate the final pathways w_{final}:

$$w_{\text{final}} = \text{LLM}_{\text{general}}\Big(\text{Prompt}(\mathcal{N}_k(e_t))\Big) \tag{9}$$

5 Experiments

In our experiment, we aim to answer the following research questions: **RQ1**: Can GRMP achieve state-of-the-art performance for mechanism pathway prediction? **RQ2**: Can GRMP generalize on unseen reactions? **RQ3**: How do the key components of GRMP contribute to its overall performance? **RQ4**: Can GRMP generate new reasonable but not explored pathways?

5.1 Experimental Setup

Datasets. We first evaluate the effectiveness of GRMP on three QA datasets. 1) **CMQA**: our primary dataset constructed directly from the CMKG. We randomly split the pathways of 50,000 reactions into training, validation, and test sets using an 8:1:1 ratio. The empirical results reported for CMQA specifically denote the performance on this held-out test set. See Appendix A, Table 8 for detailed data statistics. 2) **oMe-Gold** (Xu et al., 2025): a gold-standard benchmark, consisting of 196 literature-verified organic mechanistic pathways. 3) **CMQA-Hard-Test**: comprises 2,009 reactions drawn from eleven held-out reaction classes and 24 unseen mechanism templates, establishing a strict OOD evaluation scenario.

Baselines. We compare GRMP with the 14 baselines grouped into three categories: 1) LLM reasoning methods, 2) Mechanistic models, and 3) our KG-enhanced LLM methods. The details of baselines are listed in Appendix C.

Evaluation Metrics. Following prior work on mechanistic prediction (Joung et al., 2024, 2025), we report Validity, Main Product Accuracy (MPA), Intermediate Hit (I-hit), Pathway Accuracy (PA), and Sequence Rank (SR) metrics for evaluation. Validity measures whether generated structures are chemically well-formed; MPA and I-hit evaluate product- and intermediate-level correctness; PA measures exact match against the annotated pathway; SR assesses ranking quality and grounding in CMKG evidence, respectively. Full definitions appear in Appendix B.

Implementations. We adopt Qwen2.5-7B-Instruct as the LLM backbones, and employ a two-layer GCN (Kipf and Welling, 2016) network with BGE (Chen et al., 2024a) text embedding to learn graph representations for link prediction. We generate top-k pathways for comprehensive evaluation. During inference, we adopt the advanced Gemini-3.1-Pro for inductive reasoning. The detailed hyper-parameters and experiment settings are described in Appendix B.

5.2 Main Results (RQ1)

We evaluate the mechanism prediction performance of GRMP against state-of-the-art baselines across two benchmark datasets: CMQA and oMe-Gold. As shown in Table 1, we report the top-1 performance metrics. Notably, GRMP* represents our proposed graph-aligned LLM utilizing

Types	Methods	CMQA (%)				oMe-Gold (%)			
		Validity	MPA	I-hit	PA	Validity	MPA	I-hit	PA
LLM Reasoning	GPT-5.2 (Singh et al., 2025)	70.3	10.0	8.1	5.0	90.1	23.1	29.1	23.6
	GPT-4o (Hurst et al., 2024)	63.8	8.1	5.0	5.0	83.7	8.9	13.8	3.9
	Gemini-3.1-Pro (Team et al., 2023)	90.1	52.2	7.3	10.0	90.5	57.8	37.9	34.9
	DeepSeek-V4-Pro (DeepSeek-AI, 2026)	95.0	48.0	12.9	5.0	87.6	45.3	25.0	22.5
	GPT-5.2 + Few-shot	78.9	13.1	13.0	0.0	90.6	17.8	11.3	18.3
	GPT-5.2 + CoT (Wei et al., 2022)	73.5	38.5	7.0	0.0	**94.6**	24.7	46.9	36.9
	Llama-3.1-8B (Grattafiori et al., 2024)	40.0	0.0	0.81	0.0	73.5	12.6	3.3	2.7
	Qwen2.5-7B (Hui et al., 2024)	55.5	1.0	1.0	0.0	58.4	14.9	4.7	3.2
	ChemDFM (Zhao et al., 2025)	44.2	0.0	0.0	0.0	17.4	0.0	0.0	0.0
Mechanistic Models	Reactron (Chen et al., 2025)	93.0	**68.6**	29.3	11.8	91.8	8.6	33.6	5.2
	Graph2SMILES (Joung et al., 2024)	96.7	61.9	37.6	33.5	93.9	7.0	28.3	5.2
	FlowER (Joung et al., 2025)	**100.0**	51.8	31.4	27.2	-	-	-	-
KG-based LLM (Ours)	GRMP* (w/o general LLMs)	92.1	48.3	50.3	36.9	83.4	53.6	61.3	41.8
	GRMP (Gemini-3.1-Pro)	93.9	67.8	**64.3**	**48.1**	89.3	**73.8**	**63.6**	**57.9**

Table 1: Performance comparison with different baselines on the two mechanistic knowledge graph-based question-answer datasets (CMQA and oMe-Gold). The best results are highlighted in **bold**, and the second-best results are underlined.

Model	MPA (%)			I-hit (%)			PA (%)			SR (%)		
(CMQA)	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
FlowER	63.6	54.0	38.8	57.0	31.2	23.3	53.5	24.5	7.8	60.5	38.9	35.1
G2S	**73.4**	65.0	**48.1**	64.2	38.6	28.1	62.1	31.6	11.2	67.4	46.0	43.1
GRMP*	32.4	55.7	37.1	40.9	60.4	**41.9**	64.2	**42.0**	29.0	62.3	81.6	67.1
GRMP	57.5	**67.3**	38.1	**71.4**	**68.9**	41.1	**70.8**	41.6	**32.0**	**67.5**	**86.0**	**73.1**

Table 2: Performance comparison against mechanistic models across different reaction complexities on CMQA.

our two-stage training but without the final general-purpose LLM for inductive reasoning. The empirical results indicate that our GRMP (Gemini-3.1-Pro) consistently outperforms all baselines, achieving a Pathway Accuracy (PA) of 48.1% on CMQA and 57.9% on oMe-Gold. Specifically, our method outperforms the strongest mechanistic baseline (Graph2SMILES, 33.5% PA) by 14.6% absolute on CMQA. While general-purpose LLMs like Gemini-3.1-Pro and DeepSeek-V4-Pro exhibit high SMILES validity (up to 95.0%), their full-pathway accuracy remains poor, frequently falling at or below 10.0%. This confirms that without graph grounding, text-centric LLMs *struggle to maintain the exact structural consistency* required for multi-step pathways.

On the independent, literature-verified oMe-Gold benchmark, specialized models like FlowER and Graph2SMILES experience a performance collapse (e.g., Graph2SMILES PA drops to 5.2%). This highlights that models trained on isolated, single-step transitions fail to generalize to multi-step prediction. In contrast, our method achieves a PA of 57.9% on the oMe-Gold benchmark, outperforming the strongest standalone LLM (Gemini-3.1-Pro) by an absolute margin of 23.0%. Crucially, even without the general LLM reasoning, GRMP* alone achieves an outstanding 41.8% PA.

To further analyze the robustness of GRMP, we evaluate its performance against mechanistic models on CMQA across varying reaction complexities. Based on the number of elementary steps L in the mechanism pathways, we categorize the test set into three difficulty tiers: Easy ($L \leq 2$), Medium ($2 < L \leq 5$), and Hard ($5 < L \leq 9$). As revealed in Table 2, we conclude that: 1) traditional methods suffer from compounding errors as pathways deepen. While FlowER and Graph2SMILES perform well on easy reactions (achieving 53.5% and 62.1% PA), their performance declines sharply on hard reactions, collapsing to a meager 7.8% and 11.2% PA. 2) Pure GCN is insufficient for complex pathways. In the Appendix D.1 and Appendix D.2, pure GCN struggles as pathway depth scales, failing on hard reactions. 3) The graph-aligned LLM preserves long-range chemical logic. GRMP* retains a robust 29.0% PA and a remarkable 67.1% SR on hard reactions. 4) With Gemini-3.1-Pro

Model (CMQA -Hard-Test)	MPA (%)			I-hit (%)			PA (%)			SR (%)		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
FlowER	15.0	17.8	0.0	33.3	18.9	7.5	10.0	14.2	0.0	60.0	35.7	14.2
G2S	5.0	14.2	14.2	6.6	11.6	21.2	5.0	10.7	0.0	10.0	17.8	28.5
GRMP*	62.8	38.6	18.7	24.2	21.1	23.0	49.0	20.8	10.1	47.9	51.2	36.5
GRMP	**74.8**	**43.8**	**27.1**	**35.0**	**27.5**	**23.8**	**50.1**	**32.5**	**12.7**	**69.3**	**68.5**	**48.4**

Table 3: Performance comparison against mechanistic models across different reaction complexities on the sampled 110 reactions from the CMQA-Hard-Test set.

Variant (Predict-then-Verify)	Validity (%)			I-hit (%)			MPA (%)			PA (%)		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Mechanistic Models												
FlowER	**100.0**	**100**	**100**	18.6	19.0	20.2	14.6	14.6	14.6	10.9	10.9	10.9
G2S	89.1	90.9	90.9	15.7	17.7	18.2	12.7	14.6	14.6	7.3	9.1	9.1
Baselines & Tuning:												
SFT-only	85.4	90.6	94.5	19.9	20.7	21.0	14.5	20.0	21.0	12.7	12.7	13.6
w/ DPO ($\alpha = 0.4, \beta = 0.3$)	90.9	92.8	94.5	21.0	26.2	28.2	22.7	24.5	29.1	18.1	21.0	24.5
w/ DPO ($\alpha = 0.3, \beta = 0.3$)	93.6	93.6	96.3	23.0	25.8	29.4	35.4	36.3	38.2	27.2	36.3	36.3
Ablation Components:												
w/ RR, w/ general LLMs	89.3	96.3	96.3	**39.8**	**40.9**	**46.6**	**63.6**	**72.7**	**76.3**	**36.3**	**38.2**	**38.2**

Table 4: Ablation study evaluating the architectural components of our proposed framework across top-k configurations on the sampled 110 reactions from the CMQA-Hard-Test set. The pink marker represents the significant improvement, and the gray markers represent the slight improvements.

for inductive reasoning, our framework exhibits remarkable pathway prediction, achieving a PA of 32.0% and 73.1% SR on hard reactions. It indicates that the inductive reasoning stage successfully integrates the retrieved triples and predicted paths into a chemically consistent logic for predicting mechanism pathways. Furthermore, in Appendix D.4, context-augmented experiments on ChemCoTBench (Hao et al., 2026) show that our predicted pathways also improve the general-purpose LLM's mechanistic reasoning ability for mechanism selection.

5.3 Generalizability to Unseen Reactions (RQ2)

A persistent challenge in ML-based mechanistic modeling is the inability to generalize to OOD chemical transformations (Joung et al., 2024). We extract a subset from the CMQA-Hard-Test benchmark for evaluation, which comprises 110 reactions drawn from eleven held-out reaction classes and 24 completely unseen mechanism templates proposed by (Joung et al., 2025). Similarly, we categorize the reactions into three complexities for evaluation.

As revealed in Table 3, traditional mechanistic methods are unable to generalize to unseen reactions. FlowER achieves an MPA of 15.0% on the easy reactions but collapses entirely to 0.0% under hard ones, failing entirely to complete full-pathway predictions. In contrast, GRMP* delivers a stable performance with a 49.0% PA and a 47.9% SR on the easy subset, while outperforming baseline methods across most medium and hard reactions. Notably, the RR module drives a 10.8% absolute improvement in I-hit (from 24.2% to 35.0%), while consistently elevating MPA across all tiers (e.g., from 38.6% to 43.8% in medium). This confirms that querying the CMKG effectively anchors the model's localized chemical intuition at individual intermediate nodes. However, the improvement for PA is limited, improving from 10.1% to 12.7%. This performance bottleneck highlights the nature of multi-step mechanism generation: while the topological guidance from the RR module significantly enhances single-step and intermediate selections, a single failure in a long-horizon reasoning chain will still cause the entire pathway to fail under the strict PA metric.

5.4 Ablation Study (RQ3)

Next, we conduct an ablation study on subset datasets to evaluate the contribution of each component in the predict-then-verify framework. The

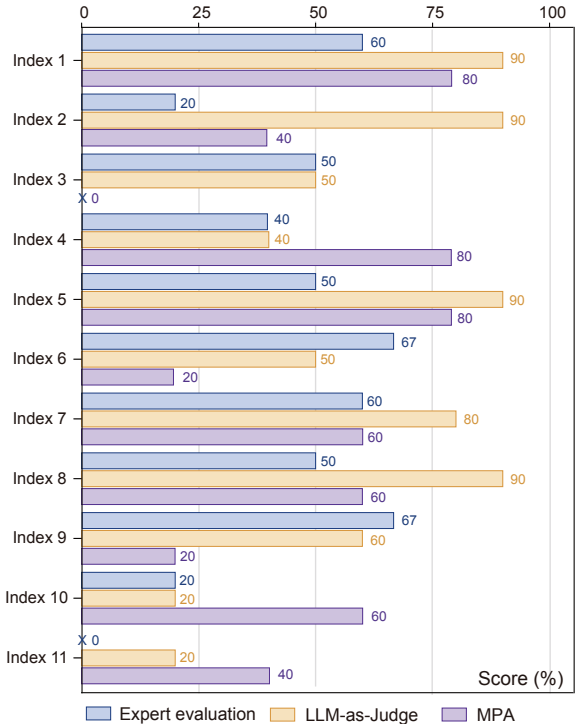

Figure 3: Results of the human evaluation and LLM-as-Judge assessment across 110 reactions on eleven OOD reaction classes.

retrieved hops for triplets in our experiments are set to three. As shown in Table 4, the SFT-only (Qwen-2.5-7B) baseline exhibits limited mechanistic reasoning ability, achieving only 14.5% Top-1 MPA and 12.7% Top-1 PA. Incorporating DPO substantially improves the pathway accuracy. Notably, the setting of hyperparameters is critical. Our optimal DPO variant ($\alpha = 0.3, \beta = 0.3$) significantly outperforms the alternative configuration ($\alpha = 0.4, \beta = 0.3$), driving the Top-1 MPA from 22.7% to 35.4% and the Top-1 PA from 18.1% to 27.2%. Crucially, I-hit is a vital metric for assessing inductive reasoning since a single chemical reaction may proceed through multiple mechanistic pathways. High I-hit and MPA demonstrate that the model is navigating a reasonable "chemical space". We find that the RR module with Gemini-3.1-Pro provides significant performance gains for I-hit, increasing Top-1 I-hit from 23.0% to 39.8%. It indicates that the RR module obtains faithful evidence from CMKG to improve the contextual accuracy of the generated pathways. We also compare the effectiveness of different general-purpose LLMs in Appendix D.3. Overall, our framework generates highly diverse and robust candidate pathways, establishing its utility as a reliable assistant for complex chemical discovery.

5.5 Human Evaluation (RQ4)

To evaluate the plausibility of our generated pathways, we conduct a human evaluation alongside an LLM-as-Judge framework on the 110 sampled OOD reactions. Predicted pathways are evaluated by three organic chemistry Ph.D. candidates using a three-tiered rubric: Reasonable (1.0), Partially Reasonable (0.5), and Unreasonable (0.0). In parallel, we prompt GPT-5.2 as an automated judge using the same expert-validated chemical rubrics (detailed in Appendix E). Finally, we systematically calculate evaluation scores across eleven distinct reaction classes, indexed from 1 to 11 as mapped in Table 6.

The results in Figure 3 show that across reaction classes where the MPA drops below 50.0%, GPT-5.2 flags four classes as chemically sound, and human experts rate six out of eleven total classes above 60.0%. This demonstrates GRMP's capacity for alternative mechanism discovery. Further, we supplement three representative case studies in Appendix F. These include: (1) a typical failure case; (2) an reaction example where experts judge the model's prediction to be more reasonable than the reference label (Figure 6); and (3) an example both the reference and the prediction are logical under the LLM judge, yet human experts prefer the model's route for its superior alignment with advanced chemical rules (Figure 7). These evaluations demonstrate that our framework can generalize beyond fixed reference templates to predict chemically plausible pathways, serving as a practical tool for novel reaction mechanism exploration.

6 Conclusion

In this work, we introduce a novel approach for reliable multi-step reaction mechanism prediction by integrating structural chemical knowledge graphs with the generative power of large language models. We a dedicated chemical mechanism KG using expert-annotated templates. We propose a predict-then-verify framework consisting of GRMP and an RR verifier for multi-step mechanism pathway reasoning. Empirical experiments across the benchmark human evaluation demonstrate the effectiveness of our approach, showcasing its potential to advance mechanism discovery.

Limitations

Despite its strong performance, our framework exhibits a few limitations that pave the way for future research. First, the retrieval efficiency hinges heavily on identifying the most relevant structural triplets from the CMKG. If the retrieval module introduces noisy triplets, it will lead to structural hallucinations for long-step mechanism reasoning. Second, expanding the current training corpus far beyond the 50,000 reactions is necessary to systematically evaluate how foundation model capabilities scale with larger mechanistic datasets. Finally, our framework still partially relies on expert-annotated mechanistic data for model development. Developing fully autonomous methods to discover and expand reaction templates without human intervention represents a vital next step for mechanism discovery.

References

Jordi Burés and Igor Larrosa. 2023. Organic reaction mechanism classification using machine learning. *Nature*, 613(7945):689–695.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Shuan Chen, Ramil Babazade, Taewan Kim, Sunkyu Han, and Yousung Jung. 2024b. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Scientific Data*, 11(1):863.

Shuan Chen, Kye Sung Park, Taewan Kim, Sunkyu Han, and Yousung Jung. 2025. Predicting chemical reaction outcomes based on electron movements using machine learning. *arXiv preprint arXiv:2503.10197*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, and 1 others. 2026. Accelerating scientific discovery with co-scientist. *Nature*, pages 1–3.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Li Hao, He Cao, Bin Feng, Daniel Shao, Robert Tang, Zhiyuan Yan, Yonghong Tian, Li Yuan, and Yu Li. 2026. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *Advances in Neural Information Processing Systems*, 38.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and 1 others. 2024. Unimol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*.

Joonyoung F Joung, Mun Hong Fong, Nicholas Casetti, Jordan P Liles, Ne S Dassanayake, and Connor W Coley. 2025. Electron flow matching for generative reaction mechanism prediction. *Nature*, 645(8079):115–123.

Joonyoung F Joung, Mun Hong Fong, Jihye Roh, Zhengkai Tu, John Bradshaw, and Connor W Coley. 2024. Reproducing reaction mechanisms with machine-learning models trained on a large-scale mechanistic dataset. *Angewandte Chemie International Edition*, 63(43):e202411296.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Daniel E Levy. 2017. *Arrow-pushing in organic chemistry: an easy approach to understanding reaction mechanisms*. John Wiley & Sons.

Haote Li, Sumon Sarkar, Wenxin Lu, Patrick O Loftus, Tianyin Qiu, Yu Shee, Abbigayle E Cuomo, John-Paul Webster, H Kelly, Vidhyadhar Manee, and 1 others. 2026. Collective intelligence for ai-assisted chemical synthesis. *Nature*, pages 1–3.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*.

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Yuan Fang Li, Chen Gong, and Shirui Pan. 2025a. Graph-constrained reasoning: Faithful reasoning on knowledge graphs with large language models. In *International Conference on Machine Learning 2025*, pages 41540–41565. Proceedings of Machine Learning Research (PMLR).

Linhao Luo, Zicheng Zhao, Gholamreza Haffari, Dinh Phung, Chen Gong, and Shirui Pan. 2025b. Gfm-rag: Graph foundation model for retrieval augmented generation. *arXiv preprint arXiv:2502.01113*.

Linhao Luo, Zicheng Zhao, Junnan Liu, Zhangchi Qiu, Junnan Dong, Serge Panev, Chen Gong, Thuy-Trang Vu, Gholamreza Haffari, Dinh Phung, and 1 others. 2025c. G-reasoner: Foundation models for unified reasoning over graph-structured knowledge. *arXiv preprint arXiv:2509.24276*.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

NextMove Software. Pistachio. https://www.nextmovesoftware.com/pistachio.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Mohammadamin Tavakoli, Pierre Baldi, Ann Marie Carlton, Yin Ting Chiu, Alexander Shmakov, and David Van Vranken. 2023a. Ai for interpretable chemistry: Predicting radical mechanistic pathways via contrastive learning. *Advances in Neural Information Processing Systems*, 36:4080–4096.

Mohammadamin Tavakoli, Yin Ting T Chiu, Pierre Baldi, Ann Marie Carlton, and David Van Vranken. 2023b. Rmechdb: A public database of elementary radical reaction steps. *Journal of chemical information and modeling*, 63(4):1114–1123.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

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

Keheng Wang, Feiyu Duan, Sirui Wang, Peiguang Li, Yunsen Xian, Chuantao Yin, Wenge Rong, and Zhang Xiong. 2023. Knowledge-driven cot: Exploring faithful reasoning in llms for knowledge-intensive question answering. *arXiv preprint arXiv:2308.13259*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ruiling Xu, Yifan Zhang, Qingyun Wang, Carl Edwards, and Heng Ji. 2025. omebench: Towards robust benchmarking of llms in organic mechanism elucidation and reasoning. *arXiv preprint arXiv:2510.07731*.

Yu Zhang, Yang Han, Shuai Chen, Ruijie Yu, Xin Zhao, Xianbin Liu, Kaipeng Zeng, Mengdi Yu, Jidong Tian, Feng Zhu, and 1 others. 2025. Large language models to accelerate organic chemistry synthesis. *Nature Machine Intelligence*, pages 1–13.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, and 1 others. 2025. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4).

A Chemical Mechanism KG (CMKG) Construction

Reaction mechanisms can be described as step-by-step sequences of elementary reactions. Understanding reaction mechanisms is fundamental to chemistry, providing mechanistic insight into chemical transformations and enabling rational reaction design. Despite the growing interest in mechanism prediction research, Table 5 highlights the key limitations of existing mechanistic datasets:

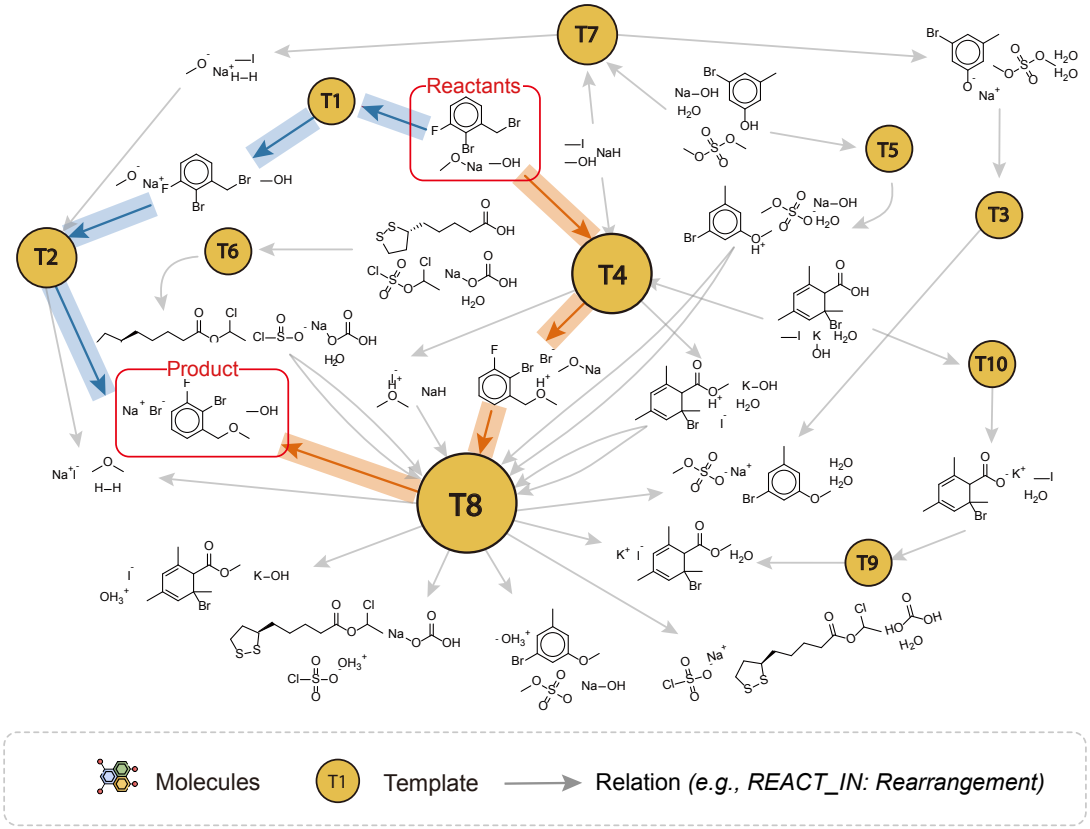

Figure 4: The subgraph visualization of our CMKG. Blue and red represent two different mechanism pathways derived from the one reaction, respectively.

the absence of complete mechanistic pathway annotations, which limits the development of multi-step mechanism prediction models. To address these limitations, we introduce a large-scale chemical mechanism dataset, a.k.a. CMQA, as shown in Figure 2. CMQA comprises 52,009 chemical reactions with complete mechanistic pathways, covering more than 180 reaction classes and 676 expert-curated elementary templates. In this section, we propose detailed information on our CMQA dataset curation.

To construct the mechanistic dataset, we leverage the Pistachio patent database (2025Q2 release), which comprises more than 15 million organic reactions. To ensure data quality, we implement a rigorous preprocessing pipeline. First, we canonicalize all reaction components (reactants, products, agents, solvents, and catalysts) in SMILES format. The complete reaction is represented as a SMILES string: "`reactants.agents.solvents.catalysts»products`". Second, building upon the methodology introduced by (Joung et al., 2024), we impute mechanistic pathways from experimentally validated reactions. After filtering samples with invalid mechanistic pathways and eliminating duplicates, we obtain 942,343 reactions and 8,222,189 elementary reactions with mechanistic pathways across 182 reaction classes and 676 reaction templates. Furthermore, we propose a two-level expert-defined taxonomy for elementary reaction steps, comprising 13 mechanism types and 259 mechanism subtypes (see in Table 6). We then prompt GPT-5.2 to generate chemical descriptions for each reaction template, and categorize the elementary transformations accordingly. All mechanistic data are manually and automatically reviewed and cleaned to enhance its consistency and reliability. Finally, from this refined pool, we randomly partition 40,000, 5,000, and 5,000 reactions for training, validation, and testing, respectively. For generalization evaluation, we additionally isolate 11 challenging reaction classes (see Table 6) to construct a hard test set, containing up to 200 reaction samples per class. This process results in the construction of the chemical mechanistic dataset, comprising 52,009 reactions and 217,441 elementary reaction steps.

After that, we construct a comprehensive chemical mechanism knowledge graph (CMKG). CMKG

Dataset	#Reactions	#Reaction Types	#Mech. Types	MI?	NL?	MP?	Cond?
Tavakoli et al. (2023b)	5k+	–	7	✓	✗	✗	✗
Chen et al. (2024b)	31k	86	13	✗	✗	✓	✗
Joung et al. (2024)	192k	86	13	✓	✗	✗	✓
Hao et al. (2026)	275	86	13	✗	✓	✗	✗
Xu et al. (2025)	2.8k	–	8	✓	✓	✓	✗
CMQA (ours)	942k	182	259	✓	✓	✓	✓

Table 5: Comparison of representative reaction mechanism datasets by scale, diversity and annotation details. "MI?", "NL", "MP", "Cond" indicate the presence of mechanistic intermediates, natural language descriptions, mechanistic pathways, and reaction conditions in SMILES format, respectively.

Category	Space
Mechanism Type	Pericyclic, Rearrangement, Radical, Electron transfer, Substitution, Cleavage, Addition, Elimination, Coordination, Proton transfer / Dissociation, Organometallic, Activation, Condensation
Mechanism Subtype	Deprotonation, Proton transfer, Elimination, Neutralization of protonated ether / Addition of alcohol under basic pKa, Nucleophilic Addition, Proton exchange, Amine addition, Halide leaves, Alkoxide leaves, Hydroxide addition, ...
Hard Reaction Type	Appel bromination (index 5), Carboxylic acid + sulfonamide condensation (index 6), Diels-Alder cycloaddition (index 9), Hydroxy to bromo (index 1), Menshutkin reaction (index 8), N-Bn deprotection (index 11), Negishi coupling (index 2), Prilezhaev epoxidation (index 10), Staudinger reduction (index 4), Transamidation (index 3), Urea Schotten-Baumann (index 7)

Table 6: Overview of mechanism and reaction types in the CMQA dataset.

Statistics	Number
Entities	247832
Molecule Nodes	247156
Template Nodes	676
Triples	422160
Reaction Types	182
Mechanism Types	13
Mechanism Subtypes	259
KG Analysis	
Max Hop	18
Average Hop	8.4432
Max Templates in-degrees	10220
Min Templates in-degrees	1
$2 \leq \text{hop} < 6$	13816
$6 \leq \text{hop} < 12$	20526
$12 \leq \text{hop} \leq 18$	15658

Table 7: Key statistics of the CMKG.

connects a wealth of chemical knowledge as a collection of triples: $\mathcal{G} = \{(e, r, e') \in \mathcal{E} \times \mathcal{R} \times \mathcal{E}\}$, where $\mathcal{E}$ and $\mathcal{R}$ denote the set of entities and relations, respectively. We define both molecular intermediates and reaction templates as entities in the CMKG. Relations describe the mechanistic roles connecting molecules and templates, including relations of the form `REACT_IN: subtype` and `PRODUCE`. Here, the mechanistic subtype specifies the elementary mechanistic transformation (e.g., "Protonation of carbonyl oxygen"). This results in 247,832 entities, of which 247,156 are molecule nodes, and 676 are template nodes. Detailed statistics are provided in Table 7. Notably, we obtain reasoning paths which are sequences of consecutive triples in the CMKG: $w_z = e_0 \xrightarrow{r_1} e_1 \xrightarrow{r_2} \ldots \xrightarrow{r_l} e_l$, where $\forall (e_{i-1}, r_i, e_i) \in \mathcal{G}$. The paths reveal the connections between knowledge that potentially facilitate reasoning.

For the application of LLM fine-tuning, we struc-

Dataset	KG	Split	Molecules	Reactions	Elem. Steps	React. Types
CMQA	CMKG	Train	197,759	40,000	168,545	167
		Validation	25,162	5,000	21,354	132
		Test	24,981	5,000	21,181	132
CMQA-Hard-Test	CMKG*	Hard-Test	8,008	2,009	6,361	11
OMeBench	-	Train	12,684	2,508	10,619	-
		Test	1,151	196	722	-

Table 8: The statistics of CMQA and OMeBench. * refers to unseen KG.

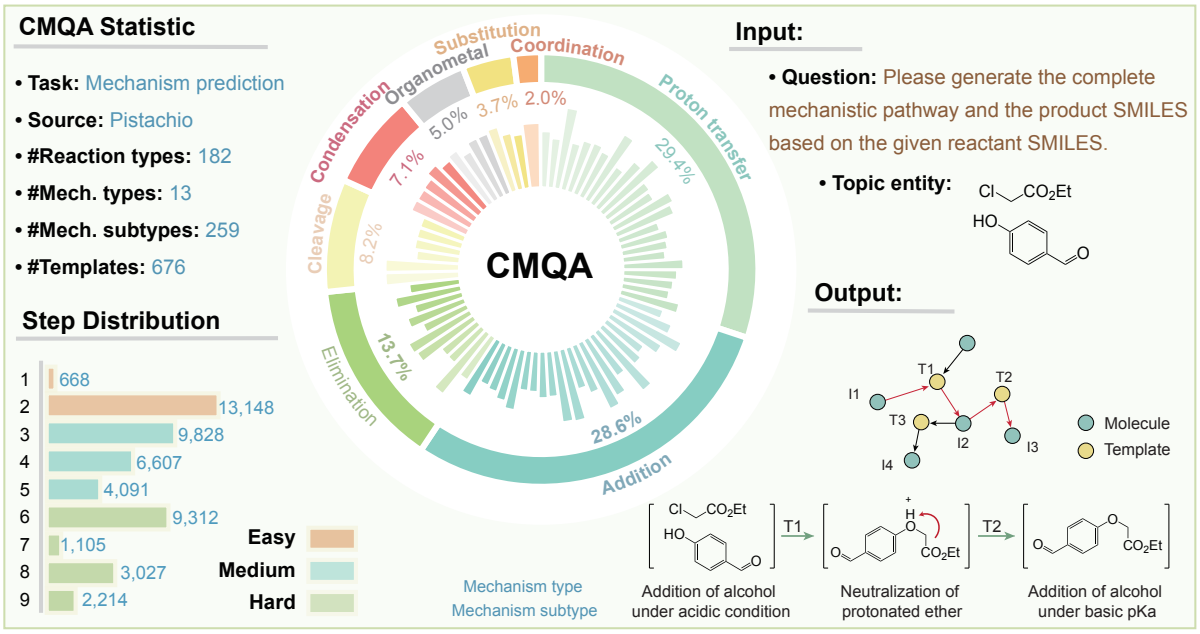

Figure 5: The demonstration of the CMQA dataset. It contains user questions and the corresponding shortest mechanism pathways as the answers.

turally integrate our CMKG into prompts and construct a pair-wise Q&A dataset, a.k.a. CMQA. To evaluate mechanistic pathway prediction, we carefully design a prompt management, which can be seen in Appendix E. Further, for each mechanistic pathway, we employ the initial reactant SMILES as the input topic entity. For instruction outputs, we constrain the generated response to the following format:

```
# Reasoning Path: <mechanistic pathway>
# Answer: <final product SMILES>
```

where `# Reasoning Path` denotes a multi-hop mechanistic reasoning trajectory in the CMKG, represented as alternating molecule-template-molecule transitions:

$$e_0 \rightarrow e_1 \rightarrow e_2 \cdots$$

Here, molecular entities correspond to reactants or intermediates. Each transition explicitly captures one elementary mechanistic transformation in the CMKG. `# Answer` denotes the final predicted product SMILES obtained at the end of the mechanistic reasoning trajectory. If multiple mechanistic pathways lead to the same product, we retain the shortest pathway. Based on the constructed CMQA dataset, we further follow the previous data splitting strategy to divide the dataset into training, validation, testing, and hard-testing sets for mechanistic pathway prediction evaluation. Detailed statistics are shown in Table 8.

B Implementation Details and Experiment Settings

In this section, we detail the implementation of GRMP and the experimental settings.

Fine-tuning specialized LLMs. We fine-tune

several foundational LLMs ranging from 0.5B to 8B. The training time and memory usage are shown in Table 10. For LLM training in both the SFT and DPO stages, we adopt LoRA-based fine-tuning. The batch size for supervised fine-tuning is set to 4, and the learning rate is set to 5e-6. The batch size for direct preference optimization training is 2, and the learning rate is 5e-7. We use the cosine learning rate scheduler policy with the warmup ratio set to 0.03. The training is conducted on four A6000-50G GPUs.

CMKG. To improve the efficiency of storage, retrieval, and reasoning, we built a chemical mechanism knowledge graph on a local Neo4j database. Neo4j is a native graph database that stores information such as nodes, relationships, and properties, making it suitable for representing complex chemical reaction networks. In CMKG, both molecules and reaction templates are treated as entities, where molecule entities include reactants, reagents, intermediates, and final products represented by `SMILES`, and template entities are expert-curated elementary reaction templates represented by `SMARTS`. For each elementary reaction step, we construct two directed triples: the input molecule is linked to its corresponding template by a REACT_IN: TYPE relation with the mechanism subtype as its label, and the template is linked to the generated molecule by a PRODUCE_IN relation. If the template is linked to the final product molecule, we define PRODUCE as a special relation. We show a practical case and a subgraph of CMKG in Table 9 and Figure 4, respectively.

GNN-based mechanism prediction. The model architecture for mechanism prediction follows an encoder–decoder for link prediction. Specifically, the encoder is a two-layer GCN, and the decoder is an MLP. For training, existing relations in the dataset are used as positive edges. Negative edges are generated by randomly double-sampling unobserved edges only between molecule nodes and template nodes. This ensures that negative samples follow the same bipartite node-type constraint as the positive edges. The encoder and decoder are optimized jointly with a binary link prediction loss. We train GNN-based predictors for 1000 epochs using Adam with a learning rate of 1×10^{-3}. The dropout rate is set to 0.5.

Metrics. Previous works (Joung et al., 2024, 2025) primarily evaluate model performance at the level of single-step elementary reactions. In this work, we focus on the more challenging task of multi-step mechanism imputation. To comprehensively assess pathway prediction quality, we adopt five evaluation metrics: Validity, Main Product Accuracy (MPA), Intermediate Hit (I-hit), Pathway Accuracy (PA), and Sequence Rank (SR). Specifically, the **Validity** metric measures the proportion of chemically valid molecules generated among the predicted mechanistic pathways. The **Main Product Accuracy (MPA)** metric evaluates whether the predicted main product exactly matches the ground-truth main product. The **Intermediate Hit (I-hit)** measures intermediate recalls by calculating the proportion of correctly predicted intermediates among all ground-truth intermediates across the dataset. **Pathway Accuracy (PA)** measures the proportion of samples whose entire predicted molecular mechanistic pathways exactly match the reference pathways, thereby evaluating end-to-end pathway reconstruction capability. Additionally, **Sequence Rank (SR)** measures whether the top-1 predicted mechanistic pathway matches any potential pathway within the CMKG. For the single-step evaluation of the GNN-based mechanism prediction, we report two metrics: First-Step Hits@k and Single-Step Hits@k. **First-Step Hits@k** evaluates whether the ground-truth template of the first elementary step in a global reaction appears among the top-k predicted templates. **Single-Step Hits@k** evaluates template prediction accuracy over all elementary steps. At each step, the prediction is considered correct if the ground-truth template is ranked within the top-k candidates given the current molecular state.

C Baseline Methods

Foundation LLMs. In this work, we evaluate a set of foundation large language models (LLMs) as baselines for chemical mechanistic reasoning, using the prompting framework provided in Appendix E. Specifically, we consider GPT-5.2 (Singh et al., 2025), Gemini-3.1-Pro (Team et al., 2023), DeepSeek-V4-Pro (DeepSeek-AI, 2026), LLaMA-3.1-8B (Grattafiori et al., 2024), Qwen-2.5-7B (Team et al., 2024), and ChemDFM (Zhao et al., 2025) as representative foundation LLMs.

- **GPT-5.2** is a frontier large language model developed by OpenAI. It is designed for complex professional tasks and provides strong capabilities in reasoning, instruction following, long-context understanding, tool use, and multimodal analysis.

$w_z =$

O=C(O[K])O[K].O=C(O)Cc1ccc(CBr)cc1.O=C(NCc1cc2c3ccccc3nc(S)n2n1)c1ccccc1OC(F)(F)F

$\xrightarrow{\text{REACT_IN: Deprotonation}}$

[S;H1;+0:1]>>[S;H0;-1:1]

$\xrightarrow{\text{PRODUCE_IN}}$

[K+].O=C(O)O[K].O=C(O)Cc1ccc(CBr)cc1.O=C(NCc1cc2c3ccccc3nc([S-])n2n1)c1ccccc1OC(F)(F)F

$\xrightarrow{\text{REACT_IN: Substitution}}$

[S;H0;-1:1].[C:2]-[F,Cl,Br,I:3]>>[S;H0;+0:1]-[C:2].[F,Cl,Br,I;-1:3]

$\xrightarrow{\text{PRODUCE}}$

[K+].[Br-].O=C(O)O[K].O=C(O)Cc1ccc(CSc2nc3ccccc3c3cc(CNC(=O)c4ccccc4OC(F)(F)F)nn23)cc1

Table 9: A case visualization in CMKG.

Model	Time	Mem. Usage
Qwen2-0.5B	3.47h	10G
Llama-3.1-8B	19.6h	23G
Qwen2.5-7B	34.5h	43G

Table 10: Training time and memory usage for different KG-specialized LLMs.

- **Gemini-3.1-Pro** is a large-scale multimodal language model developed by Google, notable for its strong reasoning performance and long-context modeling ability, which are beneficial for multi-step chemical mechanism analysis.
- **DeepSeek-V4-Pro** is the high-capacity version of the DeepSeek-V4 model family released by DeepSeek-AI. It is designed for long-context reasoning and agentic tasks.
- **Llama-3.1-8B** is an open-weight language model developed by Meta as part of the LLaMA-3.1 family. With a relatively small parameter scale, it serves as an efficient open-source baseline for evaluating whether compact general-purpose LLMs can perform chemical mechanistic reasoning under the same prompting framework.
- **Qwen-2.5-7B** is an open-source large language model developed by Alibaba, offering competitive instruction-following and reasoning capabilities at a moderate parameter scale, making it a representative efficient LLM baseline.
- **ChemDFM** is a chemistry-specific large language model fine-tuned from Llama-13B using domain-adaptive pretraining and chemistry-oriented instruction tuning, enabling enhanced chemical understanding and reasoning capabilities. In this work, we directly use the publicly available model without any architectural modification or domain-specific fine-tuning. For LLM-based models, which possess strong dialogue and instruction-following capabilities, we design task-specific prompts to enable mechanistic reasoning. The complete prompt templates are provided in Appendix E.

Mechanistic Models. In this work, we evaluate three expert models as baselines for chemical mechanistic reasoning, including Reactron (Chen et al., 2025), FlowER (Joung et al., 2025), and Graph2SMILES (Joung et al., 2024).

- **Reactron** Reactron is the first general reaction prediction model that explicitly incorporates electron movement and generates arrow-pushing mechanisms alongside product predictions.
- **FlowER** is a flow-matching-based generative model designed to model electronic rearrangement pathways in a continuous latent space. We adopt the publicly released implementation of this model and retain its default network architecture. We retrain the model on our CMQA datasets. Due to the scale difference between datasets, we slightly adjust a subset of training hyperparameters. The modified training hyperparameters are: warmup_steps = 3,000, epoch = 200, max_steps = 3,000,000, save_iter = 3,000, log_iter = 10, and eval_iter = 3,000. When

Models	First-Step		Single-Step	
	Hits@1	Hits@5	Hits@1	Hits@5
GraphSAGE	0.34	0.67	0.24	0.53
GAT	**0.57**	**0.89**	0.44	0.80
GCN	0.56	0.88	**0.50**	**0.86**

Table 11: Single-step prediction performance of different GNN architectures on the CMQA dataset.

retraining and evaluating the FlowER model, we use atom-mapped versions of the CMQA datasets, as required by the model's architecture. The same applies to Graph2SMILES. The model is trained for elementary reaction prediction.

- **Graph2SMILES** is a graph-to-sequence model that encodes molecular graphs using DGCN and decodes reaction outcomes as SMILES sequences with a Transformer. The reproduction method of Graph2SMILES is similar to that of FlowER. The modified training hyperparameters are: warmup_steps = 800, epoch = 2,000, log_iter = 10, eval_iter = 500, and save_iter = 1,000.

D Additional Experiment Results

D.1 Single-step GNN Evaluation

Table 11 reports the single-step prediction performance of different GNNs on the CMQA dataset. Given the current molecular state, the task requires the model to predict the next elementary reaction template. This task is inherently more challenging than standard single-step product prediction, as it demands a deeper internalization of reaction templates and mechanistic reasoning. For implementation, all models utilize BGE-M3 (Chen et al., 2024a) embeddings as initial node features. We evaluate three representative architectures: GraphSAGE (Hamilton et al., 2017), GAT (Veličković et al., 2017), and GCN (Kipf and Welling, 2016) across two comprehensive metrics: First-Step Hits@k and Single-Step Hits@k (with $k = 1, 5$ reported for both). Detailed definitions of these metrics are provided in Appendix B.

The empirical results yield several key observations. GAT achieves the strongest First-Step performance, obtaining a Hits@1 of 0.57 and a Hits@5 of 0.89. While GCN marginally underperforms GAT on the First-Step metric (0.56 Hits@1 and 0.88 Hits@5), it exhibits superior performance on Single-Step Accuracy, commanding the highest Hits@1 of 0.50 and Hits@5 of 0.86. We assume that Single-Step Accuracy more faithfully reflects a model's sustained capability to select the appropriate mechanistic template at each intermediate molecular transition. Therefore, we select GCN as the default backbone architecture for our GNN-based mechanism prediction.

D.2 GNN-based Mechanism Prediction Evaluation

Based on the model architecture analyzed in Appendix D.1, we report the comprehensive performance of our GCN-based mechanism prediction on the CMQA dataset in Table 12. The results show that the GCN backbone achieves highly competitive intermediate-hit rates, securing an I-hit score of 62.9% on hard reactions. This strong performance directly highlights the model's capacity to accurately formulate link predictions for consecutive intermediates within the CMKG. However, despite capturing intermediate hits, the GCN baseline suffers from a remarkably low PA of 4.5% on hard reactions when using Uni-Mol (Ji et al., 2024) as embeddings. We attribute this bottleneck to two critical limitations: (1) traditional molecular pre-trained models are predominantly optimized for stable, ground-state chemical structures. Consequently, they often fail to capture the transient electronic states or highly reactive intermediates that inherently dictate multi-step mechanistic transitions. (2) Molecular encoders cannot naturally ingest symbolic reaction templates. This limitation forces template nodes and molecular nodes into disparate, unaligned feature spaces, severely disrupting long-step link prediction.

D.3 Ablation Study on Different General-Purpose LLMs

To further investigate the effectiveness of specialized base models and general-purpose reasoning LLMs, we perform an ablation study on the GRMP (w/o DPO, w/o RR). Due to the high computational cost associated with prompting multiple frontier models, we conduct this analysis on a randomly sampled subset of 110 reactions from the CMQA-Hard-Test set. As shown in Table 13. Among the fine-tuned LLMs, Qwen2.5-7B achieves the highest PA of 12.7%, outperforming both FlowER (10.9%) and Llama-3.1-8B (10.9%). Besides, Gemini-3.1-Pro exhibits superior inductive reasoning over GPT-5.2. When paired with Qwen-2.5-

Model (CMQA)	MPA (%)			I-hit (%)			PA (%)		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
GCN (BGE emb.)	71.8	52.4	23.8	99.4	57.3	62.9	71.5	29.8	8.9
GCN (UniMol emb.)	45.0	13.0	9.9	99.2	21.1	60.9	45.0	2.2	4.5

Table 12: GCN performance on mechanism pathway prediction across different reaction complexities on the CMQA dataset.

GRMP, w/o DPO, w/o RR		**CMQA-Hard-Test**			
Base Model	General LLM	Validity	MPA	I-hit	PA
FlowER	None	100.0	14.5	18.5	10.9
Graph2SMILES	None	92.7	10.9	13.3	7.2
Llama-3.1-8B	None	66.7	8.1	11.3	10.9
	GPT-5.2	88.4	9.1	23.0	9.1
	Gemini-3.1-Pro	90.0	49.1	25.4	16.4
Qwen2.5-7B	None	85.4	14.5	19.9	12.7
	GPT-5.2	88.0	36.4	25.8	12.7
	Gemini-3.1-Pro	90.3	66.7	36.4	36.3

Table 13: Performance comparison under different KG-specialized LLMs and general LLMs on 110 samples.

Models	**Accuracy**
Gemini-2.5-pro-think	0.62
Claude3.7-sonnet-think	0.49
DeepSeek-R1	0.46
o3-mini@20250103	0.49
o1-mini@20240912	0.44
Qwen3-235B-A22B-think	0.41
Qwen3-32B-think	0.46
Llama-Nemo-49B-think	0.47
GRMP (Gemini-2.5-pro-think)	**0.83**
Improvement	0.19

Table 14: Model performance on mechanism selection in ChemCoTBench.

7B, Gemini-3.1-Pro drives the MPA to a dominant 66.7% and the full-path PA to 36.3%, outperforming the PA of 12.7% achieved by the GPT-5.2. Notably, a single chemical reaction may proceed through multiple mechanistic pathways. Consequently, I-hit is a vital metric for assessing inductive reasoning. High I-hit and MPA demonstrate that the model is navigating a reasonable "chemical space". From this perspective, the Qwen2.5-7B powered by Gemini-3.1-Pro for inductive reasoning achieves the highest 36.4% I-hit. This indicates that even in cases where the full pathway does not perfectly align with the ground truth, the model can identify the critical reactive intermediates by the capabilities of general LLMs for inductive reasoning.

D.4 Context-Augmented LLM Reasoning

To evaluate whether predicted mechanistic pathways can improve mechanistic reasoning in large language models, we conduct a context-augmented multiple-choice reasoning experiment on the ChemCoTBench dataset (Hao et al., 2026).

ChemCoTBench is designed to evaluate chemical reasoning capabilities in large language models, including molecular optimization and mechanistic understanding. In particular, we focus on its reaction mechanism selection task, where models are required to identify the most plausible mechanistic pathway from multiple candidates given reaction classes, reagents, and reaction conditions. Specifically, this task contains 275 manually curated examples adopted from (Joung et al., 2024), selected for their high quality and balanced coverage of diverse reaction mechanisms.

To integrate our structure-prior mechanistic knowledge into general-purpose LLMs, we prepend the predicted mechanistic pathway generated from reactions under identical reaction conditions as contextual reasoning guidance. The augmented prompt is formulated as *"Please answer the question based on the following mechanistic reasoning path."* followed by the generated mechanistic pathway. An example is provided in Appendix E, and the evaluation results are summarized in Table 14. As shown in the table, incorporating mechanistic guidance into the input prompt leads to a substantial improvement in performance, with Gemini-2.5-pro-think achieving a 19% increase in mechanism selection accuracy. We attribute this gain to the alignment between the generated mechanistic pathways and the underlying reaction processes described in the textual mechanism candidates. The additional mechanistic context provides chemically plausible intermediate reasoning trajectories, enabling the model to better

capture the relationships among reaction conditions, reagent roles, and elementary transformation steps. These results suggest that structure-prior mechanistic knowledge can effectively enhance the mechanistic reasoning capability of general-purpose LLMs.

E Prompts

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

Reasoning path is a sequence of triples in the KG that connects the topic entities in the question to answer entities. The Chemical KG is a reaction knowledge graph where each step follows the pattern: molecule->reaction template (SMARTS)->molecule, representing reaction mechanistic transformations. Given a question, please generate some reasoning paths in the KG starting from the topic entities to answer the question.

Question: Please generate the product SMILES based on the given reactant SMILES.

Topic entities: {Input SMILES}

Example:
{A randomly selected example}

Required output format:
Reasoning Path: molecule->reaction template SMARTS->molecule
Answer: final product SMILES
ONLY output the corrected reasoning path and final answer in the required format, and NOTHING ELSE.

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

You are an expert in understanding mechanism chemistry. Your task is to generate or predict a mechanism pathway, given the input reactants and environmental conditions.

The Chemical KG is a reaction knowledge graph in which each step follows the pattern: molecule -> reaction template (SMARTS) -> molecule, representing mechanistic transformations. Given a question, please generate reasoning paths in the KG that start from the topic entities to answer it.

Question: Please generate the product SMILES based on the given reactant SMILES.

Topic entities: {Input SMILES}

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

Reasoning path is a sequence of triples in the KG that connects the topic entities in the question to answer entities. The Chemical KG is a reaction knowledge graph where each step follows the pattern: molecule->reaction template (SMARTS)->molecule, representing reaction mechanistic transformations. Given a question, please generate some reasoning paths in the KG starting from the topic entities to answer the question.

Question: Please generate the product SMILES based on the given reactant SMILES.

Topic entities: {input SMILES}

This is my predict mechanism pathway, which contains electronic intermediates and transformations, considering SMILES validity and elementary reactions, do you think it is correct? if not, help me fix this?

Predicted pathway: {input predicted pathway}

Required output format:
Reasoning Path: molecule->reaction template SMARTS->molecule
Answer: final product SMILES

Please fix the predicted reasoning path and final answer if it is not correct. If you think the prediction is already correct, just return it without any modification. ONLY output the corrected reasoning path and final answer in the required format, and NOTHING ELSE. Do NOT output reasoning process. Do NOT leave content empty.

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

Reasoning path is a sequence of triples in the KG that connects the topic entities in the question to the answer entities. The Chemical KG is a reaction knowledge graph where each step follows the pattern: molecule->reaction template (SMARTS)->molecule, representing reaction mechanistic transformations. Given a question, please generate some reasoning paths in the KG starting from the topic entities to answer the question.

Question: Please generate the product SMILES based on the given reactant SMILES.

Topic entities: [input SMILES]

Here, I give you the predicted mechanism pathway, which contains electronic intermediates and transformations. I will also give you a set of retrieved triplets from the established chemical KG that is faithful and relevant to these predicted results:

Predicted pathway: [input predicted pathway]
Retrieved faithful triplets: [input retrieved triplets]

Considering SMILES validity, the retrieved faithful triplets, and the principle of chemical mechanisms, fix the predictions and final answer if it is not correct. If you think the prediction is already correct, just return it without any modification. The OUTPUT format is molecule->reaction template (SMARTS)->molecule in the code block, NOTHING ELSE. Do NOT output reasoning process. DO NOT leave content empty.

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

Given a reaction class, reaction conditions, and reagents represented in SMARTS format, please identify the correct sequence of elementary reaction stages based on the underlying reaction mechanism.

Reaction class: Carboxylic acid to acid chloride

Reaction conditions: $SOCl_2$ or $POCl_3$

Reagents (SMARTS format): O=[S,P](-[Cl])-[Cl].[O]=[C]-[O]

Choices:

A: Deprotonation of carboxylic acid → Addition of carbonate → Nucleophilic substitution
B: Reaction of DMF and $SOCl_2$ → Vilsmeier–Haack reagent formation → Deprotonation of carboxylic acid → Addition → Deprotonation → Nucleophilic substitution
C: Oxalyl chloride + DMF → Chloride leaves → Nucleophilic attack → Activated chloride species forms → Carboxylic acid deprotonation → Carboxylate attack → Chloride leaves → Nucleophilic attack on carbonyl → Leaving group leaves
D: Nucleophilic attack on chloride source → Chloride/Fluoride leaves → Nucleophilic attack on carbonyl → Leaving group removal → Deprotonation
E: Nucleophilic attack on chloride source → Nucleophilic attack on carbonyl → Leaving group removal + Deprotonation

Please answer the question based on the following mechanistic reasoning path.

Predicted pathway: [input reasoning path]

You may reason step by step inside `<think></think>` blocks, but only write the final choice (A, B, C, ...) inside `<answer></answer>`.

Prompt for Mechanism Pathway Prediction (General-Purpose LLMs)

Reasoning path is a sequence of triples in the KG that connects the topic entities in the question to answer entities. The Chemical KG is a reaction knowledge graph where each step follows the pattern: molecule->reaction template (SMARTS)->molecule, representing reaction mechanistic transformations. Given a question, please generate some reasoning paths in the KG starting from the topic entities to answer the question.

Question: Please generate the product SMILES based on the given reactant SMILES.

Topic entities: [input SMILES]

Here, I give you the predicted mechanism pathway, which contains electronic intermediates and transformations. I will also give you several rubrics that chemists define:

Predicted pathway: [input predicted pathway]
Rubrics:

1. Predicted mechanism pathways should follow mass conservation;
2. All mechanistic steps should remain grounded in the provided reactants;
3. The predicted final product must be chemically consistent with the provided reactants and reaction conditions;
4. The reasoning process must avoid self-reinforcing or circular electron-transfer loops;
5. The pathway must explicitly account for the influence of reaction conditions (e.g., solvent, catalyst, temperature, reagent) during mechanistic reasoning and product prediction;
6. The proposed reaction mechanism should contain a reasonable level of detail, typically within 1–10 mechanistic steps.
7. The final predicted product should not contain unpaired or artificially retained electrons.

Please make the final assessment as reasonable/partial reasonable/unreasonable for this predicted pathway following these chemical rubrics.

F Cases

Here, we supplement three representative case studies for analysis. These include: (1) a typical failure case; (2) a reaction example where experts judge the model's prediction to be more reasonable than the reference label; and (3) an example where both the reference and the prediction are logical under the LLM judge, yet human experts prefer the model's route for its superior alignment with advanced chemical rules. These evaluations demonstrate that our framework can generalize beyond fixed reference templates to predict chemically plausible pathways, serving as a practical tool for reaction mechanism exploration.

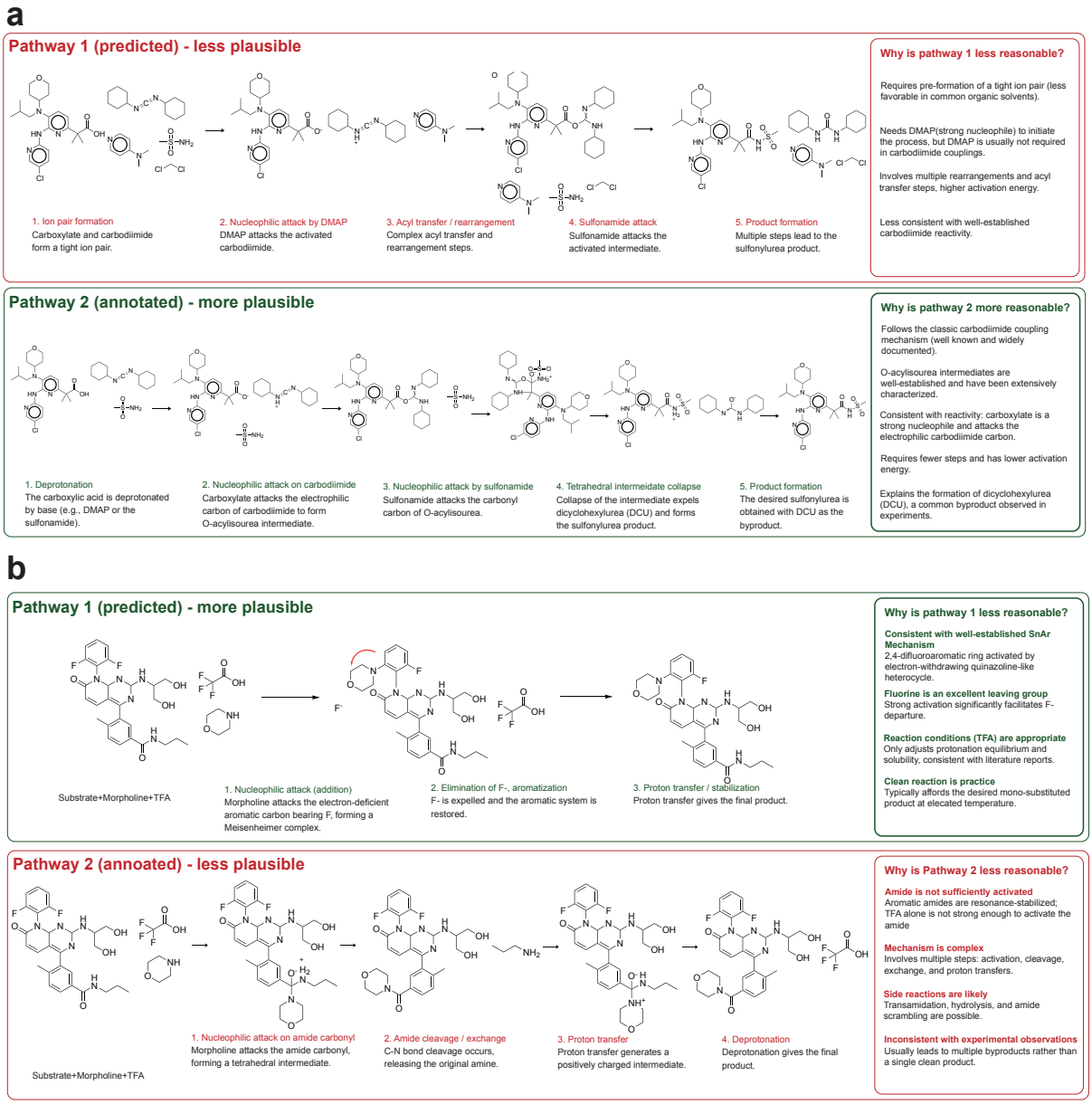

Figure 6: Two different reaction examples for multi-step mechanism prediction. **a**, A typical failure case for multi-step mechanism prediction. **b**, A reaction example where experts judge the model's prediction to be more reasonable than the reference label.

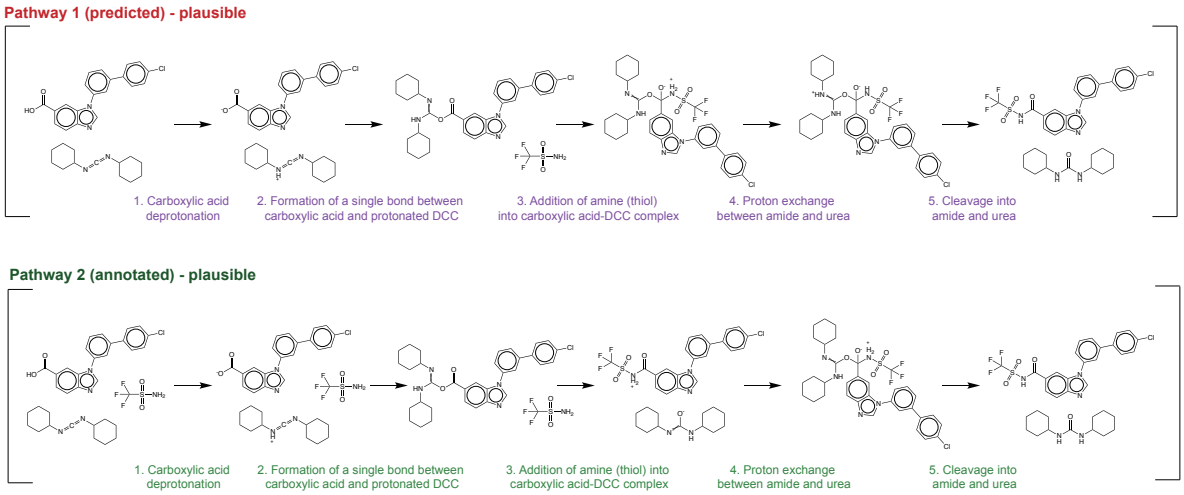

Figure 7: Predicted pathways analysis. An example where both the reference and the prediction are logical under the LLM judge, yet human experts prefer the model's route for its superior alignment with advanced chemical rules.