MPPREASONER: Reasoning-Enhanced Large Language Models for Molecular Property Prediction

Anonymous ACL submission

Abstract

Molecular property prediction is fundamental to drug discovery and materials science, yet existing models suffer from limited interpretability and poor cross-task generalization. Without explainable chemical rationales, accurate predictions alone are insufficient for guiding costly downstream experiments. We propose **MPPREASONER**, a multimodal large language model that incorporates verifiable chemical reasoning for molecular property prediction. Built on Qwen2.5-VL-7B-Instruct, MPPREASONER integrates 2D molecular images with SMILES strings through a two-stage training strategy: supervised fine-tuning (SFT) on 16,000 expert-guided reasoning trajectories, followed by Reinforcement Learning from Principle-Guided Rewards (RLPGR). RLPGR applies rule-based rewards verified by RDKit to evaluate chemical principle usage, structural analysis, and logical consistency, providing objective feedback without human preference labels. Experiments across 8 benchmark datasets show that MPPREASONER outperforms the strongest baseline by 2.95% and 5.49% on in-distribution and out-of-distribution tasks, respectively. Beyond predictive accuracy, MPPREASONER generates chemically grounded reasoning traces, with expert evaluation confirming high reasoning quality across three assessed dimensions. Code is available at `https://anonymous.4open.science/r/MPPReasoner-12687`.

1 Introduction

Molecular property prediction is a fundamental task in drug discovery and materials science, where accurate computational models can substantially reduce the cost of experimental screening (Vamathevan et al., 2019; Schlander et al., 2021; Dara et al., 2022; Sadybekov and Katritch, 2023). By estimating properties such as bioavailability, toxicity, and therapeutic efficacy before wet-lab validation, these models enable large-scale prioritization of candidate molecules from vast chemical space (Shen and Nicolaou, 2019). As molecular data become increasingly available in both structural and textual forms, there is a growing opportunity to develop models that not only make accurate predictions, but also provide interpretable evidence for decisions.

However, molecular property prediction remains challenging because molecular properties are governed by complex interactions among functional groups, substructures, physicochemical patterns, and task-specific biological mechanisms. A model may correctly classify a molecule as active or toxic, but the prediction is difficult to trust if it cannot explain which structural motifs or chemical principles support the conclusion. This issue is particularly important in scientific discovery settings, where predictions are often used to guide costly downstream experiments. In such scenarios, chemists need explanations that are chemically meaningful, logically coherent, and grounded in the actual molecular structure, rather than generic natural language rationales.

Recent advances in multimodal and language-based models have opened new possibilities for reasoning over molecular representations. Molecular images provide direct access to two-dimensional structural patterns, while SMILES strings encode atom connectivity and molecular topology in a compact textual form. These complementary representations make it possible to combine visual perception with symbolic chemical information. Nevertheless, simply prompting a general-purpose model to generate explanations does not guarantee reliable chemical reasoning. The model may overlook important functional groups, rely on superficial correlations, or produce explanations that are fluent but chemically unverifiable. Therefore, improving molecular reasoning requires not only stronger representations, but also an explicit mechanism for evaluating and optimizing the quality of reasoning.

To address this problem, we propose MPPREASONER, a multimodal framework for molecular

property prediction with verifiable chemical reasoning. MPPREASONER takes both 2D molecular images and SMILES strings as input, enabling the model to capture complementary visual and structural information. Built upon Qwen2.5-VL-7B-Instruct (Bai et al., 2025), the framework is designed to generate both property predictions and structured chemical reasoning trajectories. Instead of treating explanations as an auxiliary output, MPPREASONER explicitly models reasoning as an intermediate process that connects molecular evidence to the final prediction.

A key component of MPPREASONER is its two-stage training strategy. In the first stage, supervised fine-tuning is performed on curated chemical reasoning trajectories, allowing the model to learn how to analyze molecular structures, identify relevant substructures, and organize its reasoning in a chemically interpretable manner. In the second stage, we introduce **Reinforcement Learning from Principle-Guided Rewards (RLPGR)**, which provides verifiable, rule-based feedback for reasoning quality. The reward design evaluates whether the generated reasoning is logically consistent, grounded in molecular structure, and aligned with established chemical principles. In this way, the model is encouraged to produce explanations that are not only fluent, but also chemically faithful and useful for scientific interpretation.

Extensive experiments on 8 molecular property prediction datasets demonstrate the effectiveness of MPPREASONER. The proposed method achieves average ROC-AUC scores (Hanley and McNeil, 1982; Fawcett, 2006) of 0.8190 on in-distribution tasks and 0.7977 on out-of-distribution tasks, outperforming the best existing baselines by 2.95% and 5.49%, respectively. These results suggest that domain-grounded, verifiable reasoning rewards can contribute to both stronger predictions and improved interpretability. Expert evaluation and case studies further show that MPPREASONER generates chemically meaningful explanations, including analyses of functional groups, structural alerts, and property-related molecular patterns.

The main contributions of this work are summarized as follows:

- We propose **MPPREASONER**, a multimodal molecular property prediction framework that integrates 2D molecular images and SMILES strings to support both accurate prediction and structured chemical reasoning.
- We introduce **RLPGR**, a principle-guided reinforcement learning framework that improves reasoning quality through verifiable rewards based on logical consistency, chemical principle usage, and molecular structure grounding.
- We construct a curated dataset of 16,000 chemical reasoning trajectories to enable supervised fine-tuning of the model's chemical reasoning capabilities.
- We conduct comprehensive experiments on 8 datasets, demonstrating strong in-distribution performance, improved out-of-distribution generalization, and expert-validated reasoning quality.

2 Methodology

MPPREASONER trains multimodal LLMs to produce verified chemical reasoning through the pipeline in Figure 1. We first construct expert-level reasoning trajectories, use them for SFT, and then refine the model with RL guided by Principle-Guided Rewards.

2.1 Multimodal Molecular Prompt Design

We represent each molecule with both a 2D molecular image and its SMILES string, allowing the model to use sequential chemical tokens and visual structural cues.

As shown in Appendix G, each prompt includes a chemist role, a task description, dynamically retrieved top-5 few-shot molecules based on Morgan-fingerprint Tanimoto similarity (Bajusz et al., 2015; Cereto-Massagué et al., 2015), and the multimodal molecule input. Few-shot examples are retrieved only from the training split to avoid test leakage.

2.2 Two-Stage Training Framework

We first cold start reasoning with SFT and then refine it with RL.

2.2.1 Stage 1: Reasoning Trajectory Construction for Cold Start SFT

The first stage builds high-quality chemical reasoning trajectories through two complementary sources:

Multi-Source Reasoning Data Construction. We use strong general-domain reasoning models as teachers:

- *ChemCoT-Based One-Shot Generation:* ChemCoT exemplars (Li et al., 2025a) provide one-shot demonstrations of chemical chain-of-thought reasoning.

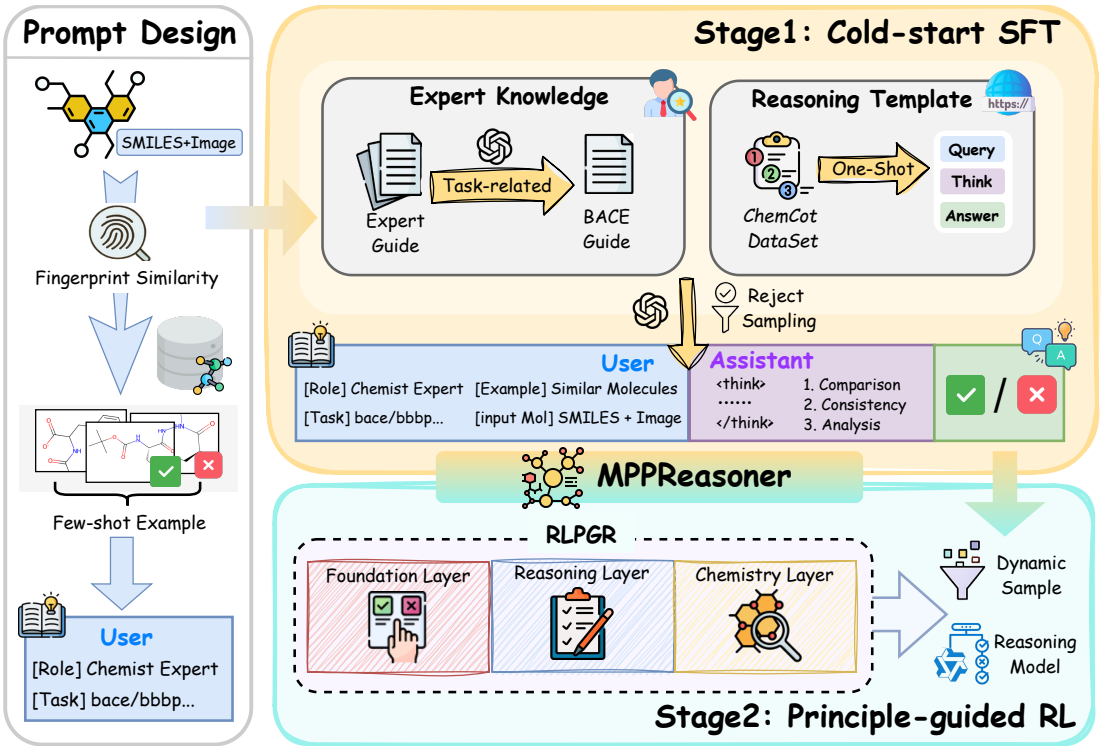


Figure 1: Overview of the MPPREASONER framework.

- *Expert-Guided Task-Specific Generation:* Five domain experts draft and cross-validate property-specific reasoning guides. GPT-4o (OpenAI et al., 2024) then extracts dataset-relevant principles (e.g., BACE and BBBP (Wu et al., 2018)) used as generation prompts.

Quality Control and Data Curation. We use rejection sampling and keep only trajectories whose teacher prediction is correct, yielding 16,000 reasoning trajectories for SFT.

We fine-tune Qwen2.5-VL-7B-Instruct with standard next-token prediction loss. GPT-4o (OpenAI et al., 2024), DeepSeek-V3.1 (Guo et al., 2025), and Qwen2.5-VL (Bai et al., 2025) generate candidate trajectories in parallel; when multiple teachers are correct, one is sampled to preserve diversity. This stage teaches the output format and instills domain-specific reasoning patterns.

2.2.2 Stage 2: Advanced Reasoning Refinement with RLPGR

While SFT teaches imitation, RLPGR improves reasoning through verifiable, rule-based rewards (Wang et al., 2025b) derived from chemical principles and RDKit computations (Landrum, 2013), avoiding dependence on human preference labels used in RLHF (Ouyang et al., 2022).

Our RLPGR framework organizes chemical reasoning feedback into three hierarchical reward layers. As illustrated in Figure 2, given a molecular description x, reasoning trace z, and prediction y, the total reward is computed as:

$$R_{\text{total}}(x, z, y) = \lambda_1 \underbrace{(r_{\text{ans}} + r_{\text{fmt}})}_{R_{\text{foundation}}} + \lambda_2 \underbrace{(r_{\text{cons}} + r_{\text{comp}})}_{R_{\text{reasoning}}} + \lambda_3 \underbrace{(r_{\text{prin}} + r_{\text{struct}})}_{R_{\text{chemistry}}} \quad (1)$$

where λ_i are hyperparameters controlling the relative importance of each reward component.

Foundation Layer: This layer ensures basic task requirements through two components:

- Answer reward r_{ans} provides binary feedback based on prediction correctness.
- Format compliance reward r_{fmt} verifies that outputs follow the required structure with reasoning enclosed in `<think>` tags and predictions in `<answer>` tags.

Reasoning Layer: This layer evaluates general reasoning quality through two key aspects:

- Logical consistency reward r_{cons} measures alignment between the reasoning conclusion and final prediction through a two-step verification process: (1) detecting conclusion keywords that indicate explicit final judgment, and (2) verifying sentiment consistency between reasoning and prediction using predefined affirmative/negative term sets for True/False predictions. Awards

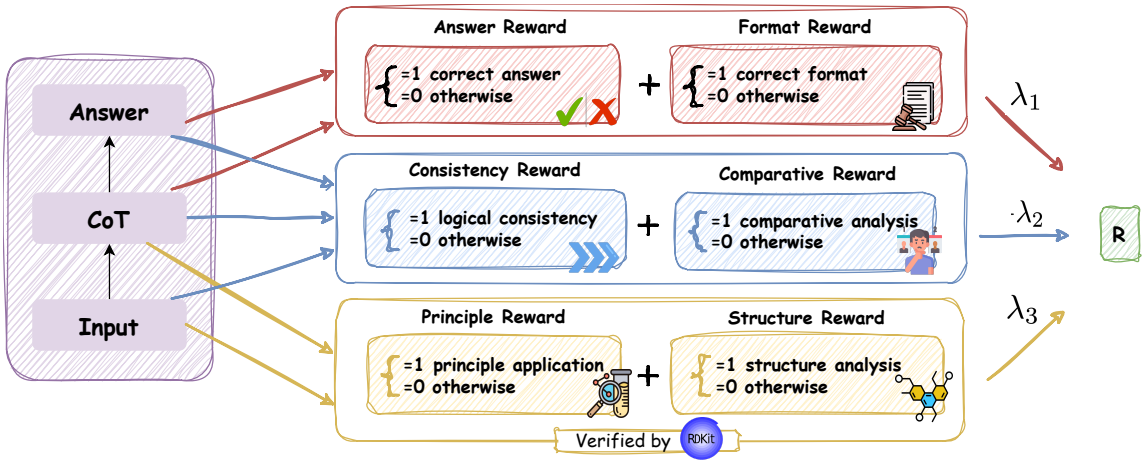


Figure 2: Illustration of RLPGR in MPPREASONER.

$r_{\text{cons}} = 1$ if both conditions are satisfied, otherwise $r_{\text{cons}} = 0$.

- Comparative analysis reward r_{comp} encourages analogical thinking by detecting explicit mentions of few-shot example SMILES strings within the reasoning text. Awards $r_{\text{comp}} = 1$ if at least one example is referenced, otherwise $r_{\text{comp}} = 0$.

Chemistry Layer: This layer targets domain-specific expertise through computational verification of chemical knowledge and structural analysis accuracy. We leverage RDKit (Landrum, 2013) for molecular property computation and substructure detection to provide objective feedback on chemical reasoning quality. The implementation uses 8 chemical-principle categories and 24 structural feature types, detailed in Appendix C.

- Chemical principle application reward r_{prin} evaluates whether mentioned chemical concepts align with computationally derived molecular properties. For instance, when the reasoning discusses hydrophobicity, we verify this against the computed LogP value. This reward ensures that chemical principles are applied appropriately rather than superficially mentioned.
- Molecular structure analysis reward r_{struct} measures the coverage of structural feature identification as:

$$r_{\text{struct}} = |S_{\text{actual}} \cap S_{\text{pred}}| / (|S_{\text{actual}}| + \epsilon)$$

where S_{actual} represents the set of distinct structural feature types identified via RDKit (functional groups, ring systems, stereochemical features), S_{pred} denotes the set of structural feature types mentioned in the reasoning trace z, and $\epsilon = 10^{-5}$ ensures numerical stability.

Training Process. We employ GRPO (Shao et al., 2024) for policy optimization, which maximizes the expected reward across reasoning trajectories:

$$\mathcal{L}_{\text{RLPGR}} = \mathbb{E}_{(x,z,y)\sim\pi_\theta}[R_{\text{total}}(x, z, y)] \quad (2)$$

where π_θ is the policy parameterized by θ. Dynamic sampling focuses training on tractable examples, encouraging the model to produce reasoning grounded in chemical principles rather than surface-level patterns.

3 Experiments

We evaluate MPPREASONER from four perspectives: overall ID/OOD performance (RQ1), the contribution of SFT and RLPGR rewards (RQ2), the value of multimodal molecular inputs (RQ3), and the quality of generated chemical reasoning (RQ4).

3.1 Experimental Setup

Datasets. We evaluate on 8 molecular property prediction datasets. ID datasets are MoleculeNet (Wu et al., 2018) benchmarks used during training: BACE (1,513), BBBP (2,039), SIDER (1,427), and HIV (41,127). We follow MoleculeNet splits: BACE, BBBP, and HIV use scaffold splits, while SIDER uses a random split. OOD datasets are held-out Therapeutic Data Commons (TDC) (Huang et al., 2021, 2022) tasks for dataset-level cross-task generalization: Bioavailability (128), CYP2C9_V (2,418), CYP2D6_V (2,626), and AMES (1,456), covering absorption,

Table 1: Performance comparison on ID and OOD benchmarks. Best is in **bold**; second best is underlined.

Model	ID Performance				OOD Performance				Average	
	BACE	BBBP	SIDER	HIV	Bioavail.	C2C9_V	C2D6_V	AMES	ID	OOD
# Molecular-domain models										
Graphormer-p	0.8575	0.7163	–	0.7788	–	–	–	–	0.7842	–
Uni-Mol	0.8570	0.7290	0.6590	**0.8080**	–	–	–	–	0.7633	–
GIMLET	0.6957	0.5939	–	0.6624	–	–	–	–	0.6507	–
MolecularGPT	0.7331	0.6822	–	0.6382	–	–	–	–	0.6845	–
Mol-LLM	0.8080	**0.8430**	0.7610	0.7650	0.6502	0.4844	0.5762	0.5469	0.7943	0.5644
InstructMol-GS	0.8210	0.7240	–	0.6890	–	–	–	–	0.7447	–
BioT5-Plus	0.8620	0.7650	0.5201	0.7630	0.5243	0.4971	0.5321	0.4466	0.7275	0.5000
MolXPT	0.8840	0.8000	0.7170	0.7810	0.4749	0.5904	0.5291	0.6073	0.7955	0.5504
3D-MoLM	0.7287	0.5141	0.5073	0.6603	0.6066	0.7081	0.7029	0.7190	0.6026	0.6842
Mol-LLaMA	0.8349	0.6263	0.5576	0.7249	0.6020	0.7556	0.7789	0.7928	0.6859	0.7323
TxGemma	0.6380	0.7102	0.5619	0.5235	0.5813	0.8075	0.7595	0.7907	0.6084	0.7348
# LLM-based generalist models										
o3-mini	0.7891	0.5972	0.5626	0.6039	0.6246	0.7729	0.7643	0.8361	0.6382	0.7495
DeepSeek-V3.1-Think	0.7017	0.6048	0.5637	0.5938	0.6572	0.7633	0.7484	0.8218	0.6160	0.7477
GPT-4o	0.6070	0.6731	0.6347	0.5698	0.5826	0.5508	0.5902	0.6141	0.6212	0.5844
Gemma-3-12b-it	0.8526	0.6802	0.5652	0.6886	0.6194	0.7886	0.7780	0.8130	0.6967	0.7498
Qwen3-8B	0.7924	0.6961	0.6442	0.5706	0.6181	0.7496	0.7165	0.8392	0.6758	0.7309
Qwen3-VL-8B-Instruct	0.8757	0.7096	0.5905	0.7143	0.5919	0.8012	0.7784	0.8514	0.7225	0.7557
Qwen3-VL-8B-Thinking	0.8597	0.7055	0.5973	0.5890	0.6073	0.7883	0.7572	0.8718	0.6879	0.7562
Qwen2.5-VL-72B-Instruct	0.7764	0.5791	0.5880	0.7325	0.6388	0.7624	0.7222	0.8156	0.6690	0.7348
Qwen2.5-VL-7B-Instruct	0.6910	0.6175	0.5823	0.5125	0.5232	0.7333	0.6999	0.7667	0.6008	0.6808
MPPREASONER (Ours)	**0.9090**	0.7436	**0.8280**	0.7932	**0.6728**	**0.8480**	**0.7950**	**0.8750**	**0.8190**	**0.7977**

metabolism, and toxicity. The OOD tasks are excluded from SFT, RL, and few-shot retrieval. We randomly sample 4,000 ID training instances with balanced task coverage and follow standard test protocols for fair comparison.

Baselines. We compare against molecular-domain models: Graphormer-p (Ying et al., 2021), Uni-Mol (Zhou et al., 2023), GIMLET (Zhao et al., 2023), MolecularGPT (Liu et al., 2024a), Mol-LLM (Lee et al., 2025), InstructMol-GS (Cao et al., 2025), BioT5-Plus (Pei et al., 2024), MolXPT (Liu et al., 2023a), 3D-MoLM (Li et al., 2024), TxGemma (Wang et al., 2025a), and Mol-LLaMA (Kim et al., 2025). We also compare against LLM-based generalist models: o3-mini (OpenAI, 2025), DeepSeek-V3.1 (Guo et al., 2025), GPT-4o (OpenAI et al., 2024), Qwen2.5-VL-72B-Instruct (Bai et al., 2025), Gemma-3-12b-it (Team et al., 2025), Qwen3-8B and Qwen3-VL-8B-Instruct/Thinking (Yang et al., 2025), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). All LLM-based baselines use the few-shot setting from Section 2.1. Implementation details are in Appendix A.

Training and Inference Details. MPPREASONER is initialized from Qwen2.5-VL-7B-Instruct and trained in two stages. The SFT stage uses 16,000 curated reasoning trajectories for 3 epochs with an effective batch size of 16, learning rate $1e{-}5$, AdamW, and 3% warmup. The RL stage runs GRPO for 500 optimization steps with 5 rollouts per input, learning rate $1e{-}6$, KL coefficient $1e{-}2$, and reward weights $(\lambda_1, \lambda_2, \lambda_3) = (1.0, 0.25, 0.25)$ for foundation, reasoning, and chemistry layers. Training uses 8 NVIDIA A100 80GB GPUs and takes about 14 hours in total. During inference, we use temperature 1.0 and top-k sampling with $k = 5$. To verify result stability, we run three independent training seeds; the standard deviations for ID and OOD average ROC-AUC are 0.020 and 0.012, both well below the reported performance gains.

3.2 Main Results (RQ1)

Table 1 shows that MPPREASONER achieves the best average performance on both ID and OOD tasks. On ID tasks, it obtains 0.8190 average ROC-AUC, outperforming the strongest baseline by 2.95%, with especially large gains on BACE and SIDER. On OOD tasks, it reaches 0.7977 average ROC-AUC, improving over the best baseline by 5.49%. These results indicate that domain-specific chemical reasoning training enables the model to achieve competitive accuracy on ID tasks while

Table 2: Ablation study on training stages and RLPGR reward.

Setting	ID Performance				OOD Performance				Average	
	BACE	BBBP	SIDER	HIV	Bioavail.	C2C9_V	C2D6_V	AMES	ID	OOD
Base (Qwen2.5-VL-7b-Instruct)	0.6910	0.6175	0.5823	0.5125	0.5232	0.7333	0.6999	0.7667	0.6008	0.6808
SFT Only	0.8558	0.6824	0.6752	0.7186	0.6625	0.7799	0.7348	0.8415	0.7330	0.7547
RL Only (RLPGR)	0.8142	0.5733	0.7428	0.5552	0.6632	0.7491	0.6732	0.7300	0.6714	0.7039
SFT + $R_{\text{foundation}}$	0.8836	0.6794	0.8089	0.7556	0.6358	0.8364	0.7862	0.8536	0.7819	0.7780
SFT + $R_{\text{foundation}}$ + $R_{\text{reasoning}}$	0.8877	0.7104	0.7981	0.7560	**0.6771**	0.8140	0.7795	0.8388	0.7881	0.7774
MPPREASONER (Ours)	**0.9090**	**0.7436**	**0.8280**	**0.7932**	0.6728	**0.8480**	**0.7950**	**0.8750**	**0.8190**	**0.7977**

Table 3: Ablation study comparing text-only and multimodal MPPREASONER.

Method	ID Performance				OOD Performance				Average	
	BACE	BBBP	SIDER	HIV	Bioavail.	C2C9_V	C2D6_V	AMES	ID	OOD
Qwen2.5-7B-Instruct	0.6950	0.5391	0.5820	0.4957	0.5094	0.7280	0.7132	0.7149	0.5780	0.6664
MPPREASONER (Text)	0.7616	0.6187	0.5599	0.5837	0.6173	0.6810	0.6771	0.7646	0.6810	0.7100
	↑9.6%	↑14.8%	↓3.8%	↑17.8%	↑21.2%	↓6.5%	↓5.1%	↑7.0%	↑17.8%	↑6.5%
Qwen2.5-VL-7B-Instruct	0.6910	0.6175	0.5823	0.5125	0.5232	0.7333	0.6999	0.7667	0.6008	0.6808
MPPREASONER (MLLM)	**0.9090**	**0.7436**	**0.8280**	**0.7932**	**0.6728**	**0.8480**	**0.7950**	**0.8750**	**0.8190**	**0.7977**
	↑31.5%	↑20.4%	↑42.2%	↑54.8%	↑28.6%	↑15.6%	↑13.6%	↑14.1%	↑36.3%	↑17.2%

retaining strong generalization: specialist models often excel only on familiar tasks, whereas generalist LLMs transfer more broadly but lack sufficient chemical grounding. Compared with the Qwen2.5-VL-7B base model, MPPREASONER improves by 36.3% on ID tasks and 17.2% on OOD tasks.

3.3 Training Strategy Ablation Studies (RQ2)

Table 2 confirms that both training stages are necessary. SFT alone improves the base model by 22.01% on ID and 10.85% on OOD tasks, showing that curated reasoning trajectories provide a strong cold start. RLPGR alone also helps, but combining SFT with RLPGR gives the largest gains (36.3% ID and 17.2% OOD). Adding reward layers progressively improves performance: foundation rewards improve task completion, reasoning rewards improve consistency and comparison, and the chemistry layer contributes the largest incremental gains (+3.92% ID and +2.61% OOD), validating computational verification of chemical principles. A no-retrieval ablation in Appendix B further shows that the gains are not mainly explained by retrieved few-shot examples.

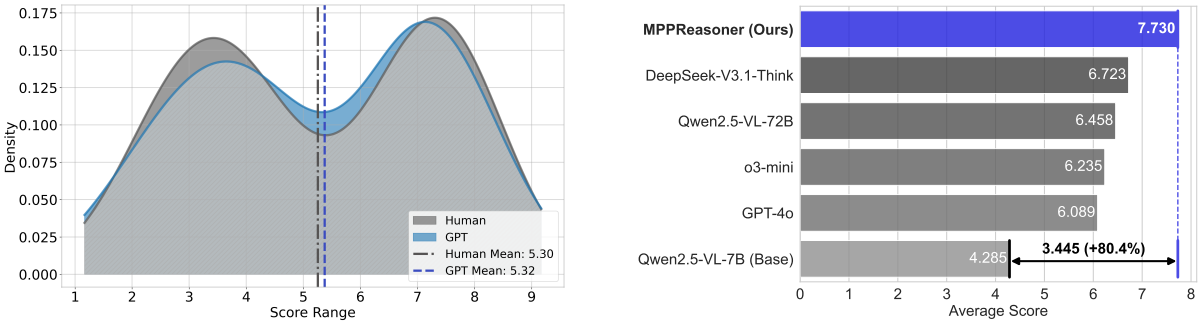
3.4 Multimodal Ablation Study (RQ3)

Table 3 compares text-only and multimodal variants trained with the same two-stage framework. The text-only model already improves over its base model, but the multimodal MPPREASONER is substantially stronger, yielding +0.1380 ID and +0.0877 OOD average ROC-AUC gains over the text-only version. This suggests that 2D molecular images complement SMILES by exposing spatial and topological cues that are difficult to recover from token sequences alone (Zeng et al., 2022; Chen et al., 2022; Feng et al., 2022).

3.5 Reasoning Quality Evaluation (RQ4)

We evaluate reasoning with a GPT-4o LLM-as-a-Judge protocol (Zheng et al., 2023; Fan et al., 2025) over logical soundness, accuracy & insight, and conciseness. To provide calibration evidence, domain experts independently score 60 samples from different datasets and models; their assessments correlate reasonably with the automated scores (Spearman $\rho = 0.82$, Figure 3), though this sample constitutes partial rather than comprehensive validation. MPPREASONER achieves the highest average reasoning score (7.730), outperforming DeepSeek-V3.1-Think (6.723), o3-mini (6.235), Qwen2.5-VL-72B (6.458), and GPT-4o (6.089). Detailed dimensional scores are in Appendix D.

Table 4 shows that the gain is not confined to one dimension. MPPREASONER improves logical soundness over the strongest baseline by 0.915 points, indicating more coherent step-by-step analysis. It also obtains the best accuracy & insight score, suggesting that RLPGR helps the model use chemical principles rather than only fluent generic explanations. The conciseness score remains competitive despite the model's longer mechanistic rea-



(a) Human-AI evaluation consistency

(b) Model reasoning quality scores

Figure 3: Reasoning quality evaluation results. (a) Automated and human assessments are consistent ($\rho = 0.82$). (b) MPPREASONER obtains the highest reasoning quality score.

Table 4: Reasoning quality scores across three evaluation dimensions. All scores are on a 0–10 scale.

Model	Logical Soundness	Accuracy & Insight	Conciseness	Average
o3-mini	7.182	5.470	6.053	6.235
DeepSeek-V3.1-Think	7.395	6.517	6.257	6.723
GPT-4o	6.698	5.916	5.653	6.089
Qwen2.5-VL-72B-Instruct	7.641	6.241	5.492	6.458
Qwen2.5-VL-7B-Instruct	4.517	3.259	5.079	4.285
MPPREASONER (Ours)	**8.556**	**7.039**	**7.352**	**7.730**

soning, which is important for expert review.

4 Case Study

To illustrate the practical benefits of our chemical reasoning approach, Figure 4 presents a representative case study comparing MPPREASONER with GPT-4o on CYP2C9 substrate prediction.

The comparison reveals fundamental differences in analytical quality: GPT-4o exhibits critical reasoning flaws including structural analysis errors (incorrectly assuming amide groups enhance binding affinity), overgeneralization (broadly claiming nitrogen-containing compounds show P450 compatibility), and incorrect reasoning patterns (unsupported statistical generalizations). These errors suggest that the generalist model relies on heuristics rather than rigorous causal analysis, leading to a wrong prediction.

In contrast, MPPREASONER follows systematic chemical reasoning: it identifies functional groups, applies CYP2C9 LogP requirements and steric effects from the bulky bicyclic indole scaffold, compares relevant structures, and keeps the final prediction consistent with the evidence. This exemplifies how RLPGR's hierarchical rewards cultivate domain-specific reasoning capabilities, enabling chemists to use the model's mechanistic insights when analyzing subtle activity cliffs. RLPGR explicitly rewards structural evidence and penalizes unsupported generalization, which is why MPPREASONER maintains reliable predictions on OOD molecules. Appendix E provides a complementary example using a bioavailability task.

5 Related Work

This section reviews prior research on machine learning for molecular representations, multimodal language models in chemistry, and reasoning capabilities in LLMs, which are foundational to our proposed framework for training reasoning LLMs tailored to molecular property prediction.

Machine Learning for Molecular Representation. GNNs have evolved as a dominant paradigm for molecular graph representation, progressing from early convolutional (Kearnes et al., 2016; Schütt et al., 2018) and message-passing (Gilmer et al., 2017) frameworks to sophisticated 3D-aware models (Gilmer et al., 2017; Li et al., 2022; Liu et al., 2023c; Zhou et al., 2023), enabling robust applications in property prediction, virtual screening, and drug discovery (Alves et al., 2022). In parallel, specialized molecular language models have reframed molecular structures as textual sequences such as SMILES strings(Weininger, 1988), with models like MolecularGPT (Liu et al., 2024a) and BioT5-Plus (Pei et al., 2024) support-

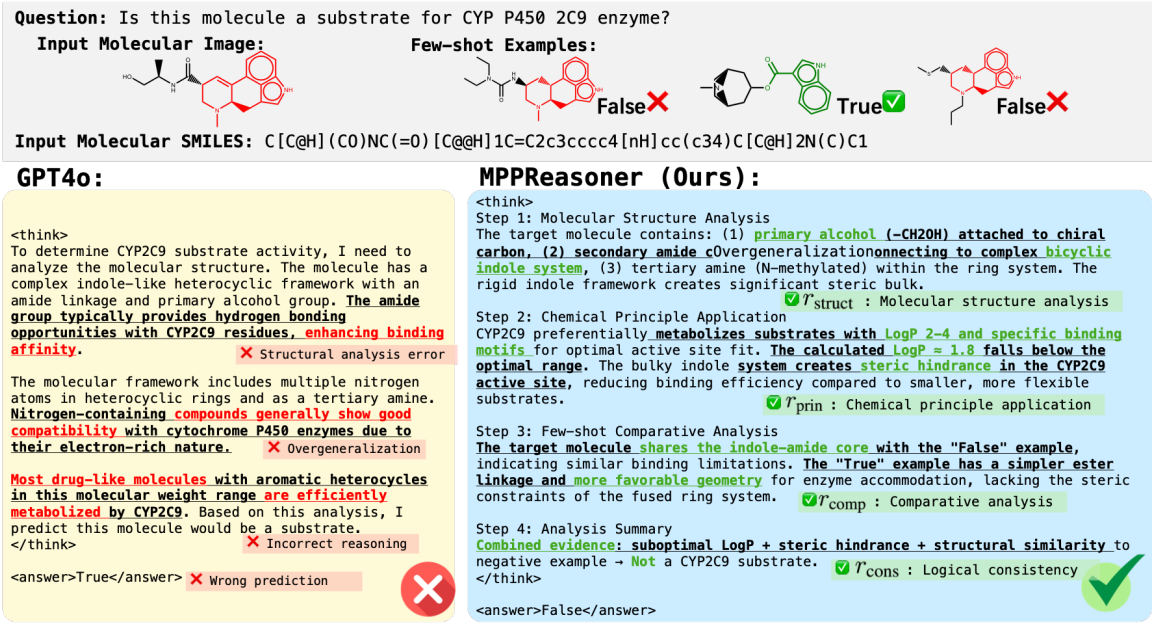


Figure 4: Case study comparison between GPT-4o and MPPREASONER for CYP2C9 substrate prediction. MPPREASONER shows systematic chemical reasoning, while GPT-4o makes several reasoning errors.

ing few-shot adaptation and multi-task learning for diverse chemical and biological tasks (Li et al., 2025c; Liu et al., 2024b, 2023b). Despite their predictive strength, these models produce point predictions without the chemical rationale needed for expert decision-making.

Language Models for Chemistry. Multimodal LLMs in chemistry integrate diverse data types such as SMILES strings and molecular graphs to address unimodal limitations, as seen in foundational molecular-text models (Lee et al., 2025; Li et al., 2025b; Liu et al., 2023b), instruction-tuned assistants (Cao et al., 2025), and tool-augmented systems (M. Bran et al., 2024). These approaches enhance robustness in property prediction (Edwards et al., 2022), molecular design (Lee et al., 2025), and synthesis planning (Shi et al., 2023; Liu et al., 2024b). Recent works (Li et al., 2024; Park et al., 2024; Kim et al., 2025; Wang et al., 2025a; Zheng et al., 2025) have enabled interpretable molecular reasoning through chain-of-thought. Together, these efforts span a broad design space including graph-text alignment (Liu et al., 2023b; Zhao et al., 2023; Cao et al., 2025), generalist molecular LLMs (Lee et al., 2025), and large-scale 3D pretraining (Zhou et al., 2023). MPPREASONER adopts SMILES strings with 2D molecular images as one practical design point in this space, and focuses on the verifiability and chemical fidelity of the generated reasoning.

Reasoning in Large Language Models. Reasoning capabilities have demonstrated remarkable efficacy in commercial LLMs, particularly through chain-of-thought (Wei et al., 2022; Jaech et al., 2024; Team et al., 2023; Priyanshu et al., 2024). Training these abilities leverages RL techniques, like Proximal Policy Optimization (Schulman et al., 2017) in RL from Human Feedback (RLHF) (Ouyang et al., 2022) for preference alignment, and efficient extensions like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) with outcome-based rewards and Reinforcement Learning with Verifiable Rewards (RLVR), improving generalization on complex tasks (Wang et al., 2025b). These RL advancements motivate our adaptation for chemical-specific reasoning in the field of molecular property prediction.

6 Conclusion

This paper presented MPPREASONER, a multimodal LLM for molecular property prediction that combines SFT with RLPGR, a verifiable reward framework for chemical reasoning. Across 8 benchmarks, MPPREASONER achieves the strongest ID and OOD performance while producing expert-validated reasoning traces. These results suggest that computationally verifiable chemical rewards can improve both predictive accuracy and interpretability, offering a practical path toward trustworthy molecular analysis.

Limitations

While MPPREASONER demonstrates significant advances in chemical reasoning for molecular property prediction, several areas present opportunities for future enhancement. First, it relies on 1D/2D molecular representations; incorporating 3D structural information (Hong et al., 2023) and conformational dynamics (Badar et al., 2022) could improve accuracy for properties sensitive to spatial arrangements. Second, the reward design and evaluation are limited to binary classification tasks; extending to regression would require redesigning the answer reward and consistency verification. Third, the evaluation covers molecular property prediction; applying the framework to reaction prediction (Chen et al., 2025b), synthesis planning (Lin et al., 2025), and molecular optimization (Kristiadi et al., 2024) remains future work.

Ethical Considerations

MPPREASONER is intended as a research assistant for molecular property prediction and should not replace experimental validation or expert safety assessment in drug discovery and materials design. Incorrect predictions or plausible but flawed explanations could mislead downstream decisions if used without human review. We therefore recommend using the model as a hypothesis-generation and interpretation aid, with final decisions supported by laboratory validation and domain experts.

Our experiments use established public molecular property benchmarks and pretrained models cited in the paper. The generated reasoning trajectories are derived from molecular inputs and expert-guided prompts rather than personal or sensitive data. The main environmental cost comes from fine-tuning and reinforcement learning on 8 NVIDIA A100 GPUs; Appendix A reports the training hardware and runtime to support reproducibility and cost assessment.

GenAI Disclosure

In the preparation of this manuscript, we utilized ChatGPT as a writing assistant. The use of the LLM was limited to proofreading for grammatical errors, checking for typos, and improving the clarity and readability of existing text. The LLM was not used for any core intellectual contributions, including but not limited to research ideation, formulation of the methodology, analysis of results, or drafting of the original manuscript. All scientific claims, arguments, and the final text are the sole work of the human authors, who take full responsibility for all content.

References

Luiz Anastacio Alves, Natiele Carla da Silva Ferreira, Victor Maricato, Anael Viana Pinto Alberto, Evellyn Araujo Dias, and Nt Jose Aguiar Coelho. 2022. Graph neural networks as a potential tool in improving virtual screening programs. *Frontiers in Chemistry*, 9:787194.

Mohammad Sufian Badar and 1 others. 2022. Molecular dynamics simulations: concept, methods, and applications. In *Transdisciplinarity*, pages 131–151. Springer International Publishing, Cham.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):20.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2025. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *COLING*.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. 2022. Structure-aware transformer for graph representation learning. In *International conference on machine learning*, pages 3469–3489. PMLR.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. 2025a. Rm-r1: Reward modeling as reasoning. *Preprint*, arXiv:2505.02387.

Zhe Chen, Zhe Fang, Wenhao Tian, Zhaoguang Long, Changzhi Sun, Yuefeng Chen, Hao Yuan, Honglin Li, and Man Lan. 2025b. Reactgpt: Understanding of chemical reactions via in-context tuning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. 2022. Machine learning in drug discovery: a review. *Artificial intelligence review*, 55(3):1947–1999.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413.

Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025. Sophiavl-r1: Reinforcing mllms reasoning with thinking reward. *Preprint*, arXiv:2505.17018.

Tom Fawcett. 2006. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Yue-Hua Feng, Shao-Wu Zhang, Qing-Qing Zhang, Chu-Han Zhang, and Jian-Yu Shi. 2022. deepmddi: A deep graph convolutional network framework for multi-label prediction of drug-drug interactions. *Analytical Biochemistry*, 646:114631.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr.

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*.

James A Hanley and Barbara J McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2021. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. 2022. Artificial intelligence foundation for therapeutic science. *Nature Chemical Biology*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Zixuan Ke, Fangkai Jiao, Yifei Ming, Xuan-Phi Nguyen, Austin Xu, Do Xuan Long, Minzhi Li, Chengwei Qin, Peifeng Wang, Silvio Savarese, Caiming Xiong, and Shafiq Joty. 2025. A survey of frontiers in llm reasoning: Inference scaling, learning to reason, and agentic systems. *Preprint*, arXiv:2504.09037.

Steven Kearnes, Kevin McCloskey, Marc Berndl, and 1 others. 2016. Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8):595–608.

Dongki Kim, Wonbin Lee, and Sung Ju Hwang. 2025. Mol-llama: Towards general understanding of molecules in large molecular language model. *arXiv preprint arXiv:2502.13449*.

Agustinus Kristiadi, Felix Strieth-Kalthoff, Marta Skreta, Pascal Poupart, Alán Aspuru-Guzik, and Geoff Pleiss. 2024. A sober look at llms for material discovery: Are they actually good for bayesian optimization over molecules? *Preprint*, arXiv:2402.05015.

Greg Landrum. 2013. Rdkit documentation. *Release*, 1(1-79):4.

Chanhui Lee, Hanbum Ko, Yuheon Song, YongJun Jeong, Rodrigo Hormazabal, Sehui Han, Kyunghoon Bae, Sungbin Lim, and Sungwoong Kim. 2025. Mol-llm: Multimodal generalist molecular llm with improved graph utilization. *arXiv preprint arXiv:2502.02810*.

Hao Li, He Cao, Bin Feng, Yanjun Shao, Xiangru Tang, Zhiyuan Yan, Li Yuan, Yonghong Tian, and Yu Li. 2025a. Beyond chemical qa: Evaluating llm's chemical reasoning with modular chemical operations. *arXiv preprint arXiv:2505.21318*.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, and 1 others. 2025b. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 415–423.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2025c. Scilitllm: How to adapt LLMs for scientific literature understanding. In *The Thirteenth International Conference on Learning Representations*.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. Towards 3d molecule-text interpretation in language models. *arXiv preprint arXiv:2401.13923*.

Sihang Li, Xiang Wang, An Zhang, Yingxin Wu, Xiangnan He, and Tat-Seng Chua. 2022. Let invariant rationale discovery inspire graph contrastive learning. In *International conference on machine learning*, pages 13052–13065. PMLR.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. Holistic evaluation of language models. *Preprint*, arXiv:2211.09110.

Xuan Lin, Qingrui Liu, Hongxin Xiang, Daojian Zeng, and Xiangxiang Zeng. 2025. Enhancing chemical reaction and retrosynthesis prediction with large language model and dual-task learning. *Preprint*, arXiv:2505.02639.

Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. 2024a. Moleculargpt: Open large language model (llm) for few-shot molecular property prediction. *arXiv preprint arXiv:2406.12950*.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023a. Molxpt: Wrapping molecules with text for generative pre-training. *Preprint*, arXiv:2305.10688.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024b. Reactxt: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5353–5377.

Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. Rethinking tokenizer and decoder in masked graph modeling for molecules. *Advances in Neural Information Processing Systems*, 36:25854–25875.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. *Preprint*, arXiv:1710.03740.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

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.

Jinyoung Park, Minseong Bae, Dohwan Ko, and Hyunwoo J Kim. 2024. Llamo: Large language model-based molecular graph assistant. *Advances in Neural Information Processing Systems*, 37:131972–132000.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. Biot5+: Towards generalized biological understanding with iupac integration and multi-task tuning. *arXiv preprint arXiv:2402.17810*.

Aman Priyanshu, Yash Maurya, and Zuofei Hong. 2024. Ai governance and accountability: An analysis of anthropic's claude. *arXiv preprint arXiv:2407.01557*.

Anastasiia V Sadybekov and Vsevolod Katritch. 2023. Computational approaches streamlining drug discovery. *Nature*, 616(7958):673–685.

Michael Schlander, Karla Hernandez-Villafuerte, Chih-Yuan Cheng, Jorge Mestre-Ferrandiz, and Michael Baumann. 2021. How much does it cost to research and develop a new drug? a systematic review and assessment. *Pharmacoeconomics*, 39(11):1243–1269.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Kristof Schütt, Huziel Sauceda, Pieter-Jan Kindermans, Stefan Chmiela, Klaus-Robert Müller, and Alexandre Tkatchenko. 2018. Schnet-a deep learning architecture for molecules and materials. *Bulletin of the American Physical Society*, 63.

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

Jie Shen and Christos A Nicolaou. 2019. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32:29–36.

Yaorui Shi, An Zhang, Enzhi Zhang, Zhiyuan Liu, and Xiang Wang. 2023. Relm: Leveraging language models for enhanced chemical reaction prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5506–5520.

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*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, and 1 others. 2019. Applications of machine learning in drug discovery and development. *Nature reviews Drug discovery*, 18(6):463–477.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. 2025a. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025b. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, pages 24824–24837.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhuohao Yu, Jiali Zeng, Weizheng Gu, Yidong Wang, Jindong Wang, Fandong Meng, Jie Zhou, Yue Zhang, Shikun Zhang, and Wei Ye. 2025b. Rewardanything: Generalizable principle-following reward models. *Preprint*, arXiv:2506.03637.

Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. 2022. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016.

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2023. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. In *Advances in Neural Information Processing Systems*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

A Implementation Details

We implement MPPREASONER based on Qwen2.5-VL-7B-Instruct (Bai et al., 2025), configured with a maximum sequence length of 8,192 tokens to accommodate detailed reasoning outputs. Our implementation follows a two-stage training pipeline with carefully tuned hyperparameters for optimal performance.

SFT stage employs 16,000 curated reasoning trajectories over 3 epochs. We use an effective batch size of 16 with a learning rate of 1e-5 and the AdamW optimizer. A linear learning rate scheduler with 3% warmup ratio ensures stable training convergence.

RL stage utilizes the GRPO algorithm (Shao et al., 2024) for 500 optimization steps with dynamic sampling (Yu et al., 2025a) to filter training instances and focus on tractable reasoning examples. SFT and RL use the same molecular inputs: SFT learns from fixed teacher-generated trajectories, while RL samples new reasoning traces for those inputs and scores them with RLPGR. Dynamic sampling removes rollout groups that are entirely correct or entirely incorrect, focusing updates on informative cases. We employ a lower learning rate of 1e-6 with weight decay of 1e-2 and KL coefficient of 1e-2 to maintain stability during policy optimization. The rollout configuration generates 5 samples per input with temperature 1.0, using a global batch size of 128 and rollout batch size of 512 for efficient training. The hierarchical reward weights in RLPGR are set as (λ_1, λ_2, λ_3) = (1.0, 0.25, 0.25) for foundation, reasoning, and chemistry layers respectively.

All training is conducted on 8 NVIDIA A100 80GB GPUs with mixed precision (Micikevicius et al., 2018) training for memory efficiency. The SFT stage requires approximately 2 hours, while the RL stage takes 12 hours, totaling 14 hours for complete training. During inference, we use temperature 1.0 with top-k sampling (k=5) to generate diverse yet high-quality reasoning paths.

B Evaluation Protocol and Additional Metrics

For ID evaluation, we follow MoleculeNet benchmark splits: BACE, BBBP, and HIV use scaffold splits, while SIDER uses the standard random split. For dataset-level OOD evaluation, Bioavailability, CYP2C9_V, CYP2D6_V, and AMES are selected because they cover absorption, metabolism, and toxicity while preserving the binary classification format used by our training objective. All LLM-based baselines use the same prompt template, top-5 training-set retrieval by Morgan-fingerprint Tanimoto similarity, temperature 1.0, and maximum generation length of 8,192 tokens.

Table 5: Hyperparameters Setting

Hyperparameter	**Value**
# Supervised Fine-tuning (SFT)	
GPU Number (A100)	8
Train Batch Size Per Device	2
Gradient Accumulation Steps	4
Learning Rate	1.0e-5
Number of Train Epochs	3
LR Scheduler Type	Linear
Warmup Ratio	0.03
# Reinforcement Learning (RL)	
GPU Number (A100)	8
Learning Rate	1.0e-6
Weight Decay	1.0e-2
KL Coefficient	1.0e-2
Rollout Number	5
Rollout Temperature	1.0
Global Batch Size	128
Rollout Batch Size	512
Micro Batch Size For Update Per Device	8
# Inference	
Temperature	1.0
Top K	5
Max Tokens	8,192

Table 6: No-retrieval ablation. Few-shot examples are removed from both training and inference; the comparative reward is disabled accordingly.

Setting	**ID Avg.**	**OOD Avg.**
Qwen2.5-VL-7B-Instruct (Zero-shot)	0.5155	0.5157
Qwen2.5-VL-7B-Instruct (Few-shot)	0.6008	0.6808
MPPREASONER (Zero-shot)	0.7813	0.7794
MPPREASONER (Few-shot)	**0.8190**	**0.7977**

Table 6 shows that MPPREASONER remains strong without retrieved examples, substantially outperforming the few-shot base model. This indicates that the gains are not mainly attributable to nearest-neighbor prompting, although retrieval provides additional improvement on top of the trained model.

Table 7: F1 scores for representative LLM-based methods. ROC-AUC is the primary metric following MoleculeNet convention, while F1 shows the same trend under a 0.5 decision threshold.

Model	ID Avg.	OOD Avg.
Qwen3-VL-8B-Thinking	0.6022	0.6747
Qwen2.5-VL-7B-Instruct	0.5516	0.5838
MPPREASONER (Ours)	**0.6559**	**0.7053**

C RLPGR Rule Details

The chemistry layer is strictly rule-based and uses RDKit-derived molecular properties or substructure matching. The chemical principle reward r_{prin} covers 8 categories: Lipinski's rule, LogP/lipophilicity, hydrogen bonding, molecular weight, aromaticity, polarity, steric hindrance, and rotatable bonds. For example, hydrophobicity mentions are checked against computed LogP, Lipinski claims are checked against molecular weight, LogP, hydrogen-bond donors, and hydrogen-bond acceptors, and flexibility claims are checked against the number of rotatable bonds.

The structure reward r_{struct} covers 24 structural feature types: hydroxyl, carboxyl, amino, nitro, ester, amide, ether, ketone, aldehyde, cyano, benzene, pyridine, indole, quinoline, naphthalene, cyclohexane, cyclopentane, heterocycle, fluorine, chlorine, bromine, chiral centers, branched chains, and linear chains. These features are verified through RDKit fragment functions, SMARTS matching, atom inspection, ring inspection, or chiral-center detection. No learned or neural verifier is used in the reward computation.

D Reasoning Quality Scores

Table 4 in the main text presents the detailed reasoning quality scores across three evaluation dimensions (Fan et al., 2025) for all models in our study. The detailed dimensional analysis reveals several important insights into model capabilities and reasoning patterns (Ke et al., 2025; Liang et al., 2023; Yu et al., 2025b; Chen et al., 2025a). MPPREASONER achieves the highest scores across all three evaluation dimensions, demonstrating comprehensive reasoning excellence. In logical soundness, MPPREASONER scores 8.556, significantly outperforming the best baseline DeepSeek-V3.1-Think at 7.395, indicating superior coherence in step-by-step reasoning flow. For accuracy & insight, our model achieves 7.039, substantially exceeding DeepSeek-V3.1-Think's 6.517, which demonstrates the effectiveness of chemical principle integration in generating factually correct and insightful analyses.

Examining model category patterns, advanced reasoning models like o3-mini and DeepSeek-V3.1-Think show relatively strong logical soundness but struggle with accuracy & insight, particularly o3-mini at 5.470, suggesting that general reasoning capabilities cannot substitute for domain-specific knowledge. Large-scale models exhibit mixed performance: Qwen2.5-VL-72B-Instruct achieves decent logical soundness (7.641) but suffers in conciseness (5.492), while the smaller Qwen2.5-VL-7B-Instruct shows consistently poor performance across all dimensions, with particularly low accuracy & insight at 3.259. Notably, MPPREASONER maintains balanced excellence across all dimensions, avoiding the trade-offs observed in baseline models. The model's conciseness score of 7.352 is particularly remarkable, as it demonstrates the ability to provide comprehensive chemical reasoning without unnecessary verbosity, a crucial factor for practical applications where chemists need clear and actionable insights.

E Comparison Study: Text-Only vs. Multimodal

To illustrate how visual information contributes to reasoning quality, we provide a comparative study analyzing the same oral bioavailability prediction task using both text-only and multimodal versions of MPPREASONER. The input molecule, shown in Figure 7, has ground truth label False.

Figure 7 shows that MPPREASONER (MLLM) correctly predicts False, identifying the problematic nitrophenyl group and its negative impact on membrane permeability. In contrast, MPPREASONER (Text-Only) incorrectly predicts True (Figure 8), despite following a seemingly logical reasoning process.

The text-only model exhibits two critical failure patterns:

- **Functional Group Misidentification.** The SMILES pattern `[N+](=O)[O-]` represents a nitro group—a strongly polar, electron-withdrawing substituent that reduces membrane permeability. However, the text-only model misinterprets this as an "oxime-like" or "oxime-type" N–O fragment, incorrectly attributing favorable

Table 8: Quality assessment of SFT reasoning trajectories across 16,000 training samples. Scores are on a 0–10 scale evaluated by GPT-4o following the rubric in Appendix G.

Tasks	Logical Soundness	Accuracy & Insight	Conciseness	Average
BACE (907)	8.17	6.69	7.14	7.33
BBBP (1169)	8.67	8.17	8.31	8.39
SIDER (5714)	7.91	7.00	7.04	7.31
HIV (8210)	8.46	7.30	7.38	7.71
Overall (16000)	**8.26**	**7.22**	**7.31**	**7.59**

hydrogen-bonding properties. This reflects a fundamental weakness: SMILES-only models rely on token co-occurrence patterns for functional group recognition rather than structural chemistry rules. The visual representation immediately reveals the nitro group's geometry and aromatic attachment, enabling correct identification.

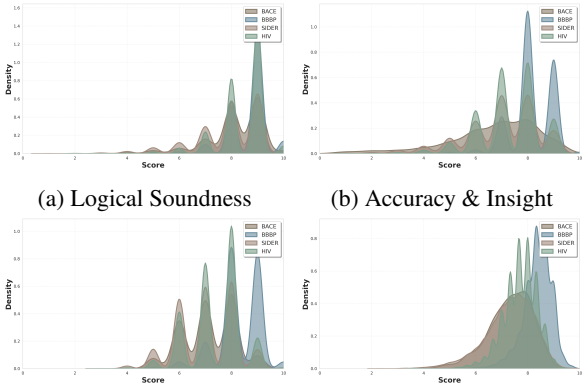
- **Semantic Prior Override.** Rather than reasoning from structural features and few-shot patterns, the text-only model anchors on semantic similarity to known drugs. It categorizes the molecule as "Dapoxetine-like oral agents" and invokes the generalization that "such scaffolds are widely reported... as orally bioavailable drugs." Once committed to this "drug-like scaffold" framing, the model defaults to predicting good bioavailability despite the nitrophenyl liability. The multimodal model, by contrast, correctly identifies the nitrophenyl group through visual inspection and recognizes its structural similarity to negative few-shot examples.

This case study demonstrates two key advantages of 2D molecular images: (1) accurate functional group recognition through direct visual perception, and (2) structural-detail-first reasoning that resists override by high-level semantic priors. These capabilities are essential for molecular property prediction requiring precise chemical analysis.

F SFT Data Quality Assessment

To validate the quality of our supervised fine-tuning (SFT) data, we conducted a comprehensive quality assessment using the same LLM-as-Judge evaluation framework described in Section 3.5. All 16,000 reasoning trajectories in our SFT dataset were evaluated across three dimensions: Logical Soundness, Accuracy & Insight, and Conciseness. Table 8 presents the average scores for each task.

The results demonstrate consistently high-quality reasoning across all tasks, with an overall average score of 7.59. Notably, Logical Soundness achieves the highest score (8.26), indicating that our multi-source generation strategy combined with rejection sampling effectively produces logically coherent reasoning chains. The Accuracy & Insight dimension scores 7.22, reflecting the integration of expert-curated chemical knowledge into the reasoning process. Conciseness scores 7.31, showing that our trajectories maintain focused and relevant explanations without excessive verbosity.



(a) Logical Soundness (b) Accuracy & Insight

(c) Conciseness (d) Average

Figure 5: Score distribution of SFT reasoning trajectories across three evaluation dimensions.

Figure 5 presents the score distribution across all three evaluation dimensions. The distributions show strong concentration in the 7-9 range, with minimal low-quality samples (scores below 5), confirming that our quality control mechanisms—including rejection sampling, expert guidance, and multi-source generation—effectively filter out poorly reasoned trajectories.

G Prompt Templates

Example 1: Prompt Example for BACE Task

[Role]
You are a top AI assistant specializing in molecular chemistry and drug discovery, proficient in molecular property prediction.

[Task]
BACE1 is an aspartic-acid protease important in the pathogenesis of Alzheimer's disease, and in the formation of myelin sheaths...
output "True" or "False".

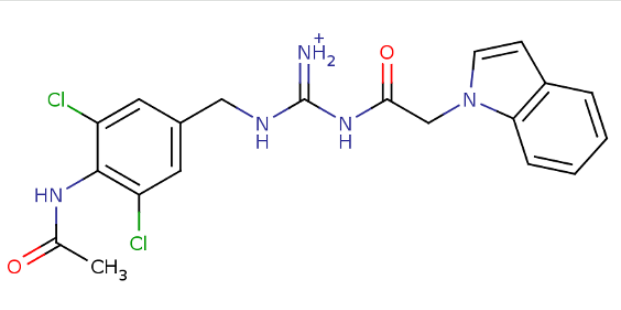
[Formatting]
Place the thought process within <think></think> and then conclude your answer with <answer>True/False</answer>.

[Example]
<think>xxxx</think>
<answer>True/False</answer>

[Few-shot]

ClC1=CC(=CC(Cl)=C1NC(=O)C)CNC(=[NH2+1])NC(=O)CN2C3=CC(OC)=CC=C3C=C2	False
ClC1=CC(=CC(Cl)=C1NC(=O)C)CNC(=[NH2+1])NC(=O)CN2C3=CC(CC)=CC=C3C=C2	True
ClC1=CC(=CC(Cl)=C1NC(=O)C)CNC(=[NH2+1])NC(=O)CN2C3=CC(F)=CC3C=C2	False

[Molecule]
ClC1=CC(=CC(Cl)=C1NC(=O)C)CNC(=[NH2+1])NC(=O)CN2C3=C(C=CC=C3)C=C2



Example 2: Prompt for ChemCoT-Based One-Shot Generation

Example Prompt:
< PROMPT retrieved from OpenMol/ChemCoTDataset >

Example Response:
< RESPONSE retrieved from OpenMol/ChemCoTDataset >

Prompt:
< PROMPT similar to Example 1 >

Response:

Example 3: Prompt for Expert-Guided Task-Specific Generation

< PROMPT similar to Example 1 >

[Expert]

< EXPERT KNOWLEDGE refined by GPT-4o >

Example 4: Prompt for Logical Soundness Scoring

You are a professional reasoning-evaluation expert. Your task is to assess the **logical soundness** of a large language model's chain-of-thought when answering a question, and assign an integer score from **0 to 10**. Focus strictly on the logical connections between reasoning steps, not on whether the final answer is correct.

Input:
- [Question]: The original question.
- [Model Response]: The model's full response, including its chain of thought.

Scoring Dimension (Logical Soundness):
- Do reasoning steps build progressively and refer back to earlier points?
- Is each step a reasonable extension of the previous inference?
- Is the language coherent, with no contradictions or confusing wording?

Scoring Scale (0-10):
- **10**: Perfect logical structure; steps are crystal-clear and fully justified.
- **8-9**: Overall logic sound; only minor or negligible leaps/wording issues.
- **6-7**: Main logic correct, but some jumps, insufficient explanation, or minor conflicts.
- **4-5**: Noticeable breaks or missing key inferences, yet some coherent logic remains.
- **2-3**: Most steps lack causality or contradict each other; only sporadic correct parts.
- **0-1**: Virtually no discernible valid reasoning structure.

Your Task:
Adhering strictly to the rubric above, you must output only a single integer score from 0 to 10. Do not provide any additional explanations, text, or justifications.

Question:
< QUESTION >

Model Response:
< RESPONSE >

Output Format: [integer score]

Example 5: Prompt for Accuracy & Insight Scoring

You are a professional reasoning-evaluation expert. Your task is to assess the **accuracy and insight value** of a large language model's chain-of-thought when answering a question, and assign an integer score from **0 to 10**...

Scoring Dimension (Accuracy & Insight):
- Are the concepts, formulas, and facts used accurate and appropriate?
- Do the reasoning perspective, decomposition approach, or intermediate conclusions provide substantive support or fresh insights for domain experts?

Scoring Scale (0-10):
- **10**: All methods and facts are completely correct, offering deep and original insights.
- **8-9**: Core content is correct, with only minor detail errors or slightly shallower insights.
- **6-7**: Mostly correct, but with notable secondary errors or average insight depth.
- **4-5**: Mix of correct and incorrect information; limited insight value.
- **2-3**: Most methods/facts are wrong or misused, providing almost no insight.
- **0-1**: Completely incorrect or irrelevant.

...

Example 6: Prompt for Conciseness Scoring

You are a professional reasoning-evaluation expert. Your task is to assess the **conciseness** of a large language model's chain-of-thought when answering a question, and assign an integer score from **0 to 10**...

Scoring Dimension (Conciseness):
- Does the response go straight to the point, avoiding irrelevant or repetitive explanations?
- Does it convey the full reasoning with the minimum necessary steps?

Scoring Scale (0-10):
- **10**: Extremely concise, with no redundant or repetitive statements.
- **8-9**: Generally concise, with only a tiny amount of removable content.
- **6-7**: Noticeable redundant paragraphs or repeated explanations.
- **4-5**: Long-winded and repetitive; key information diluted by noise.
- **2-3**: Large portions are irrelevant or repetitive; core points hard to discern.
- **0-1**: Almost entirely made up of redundant content.

...

H More Cases

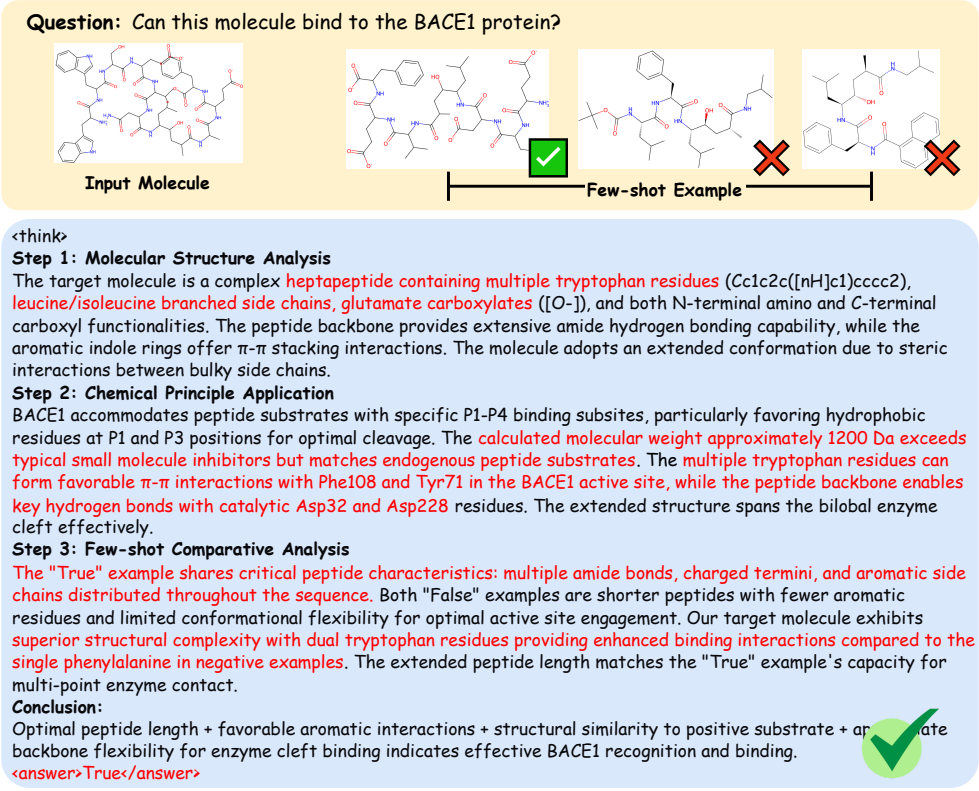


Figure 6: Successful case on MPPREASONER for BACE1 protein binding prediction (ID).

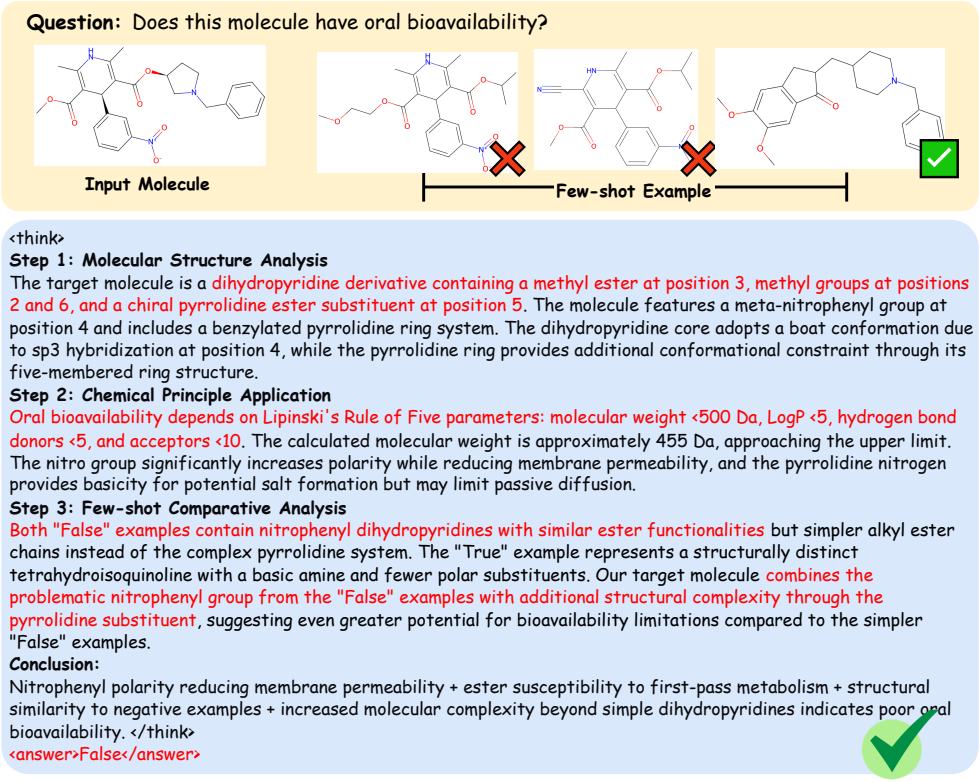


Figure 7: Successful case on MPPREASONER for oral bioavailability prediction (OOD).

<think>
Step 1: Parse the SMILES and identify key fragments
The SMILES`COC(=O)C1=C(C)NC(C)=C(C(=O)O[C@H]2CCN(Cc3ccccc3)C2)[C@H]1c1cccc([N+](=O)[O-])c1`
describes a dihydropyridine-like core with two ester groups and a substituted aromatic ring.On the aromatic ring I see a strongly polar, **oxime-like moiety `[N+](=O)[O-]`**. I interpret this pattern as an "oxime-type" N–O fragment that can contribute additional polarity and hydrogen-bonding capacity, similar to other oxime/heteroatom-rich groups. Together with the carbonyl oxygens in the esters, this gives several potential hydrogen-bond acceptors, and the "oxime-like" unit may also behave as a hydrogen-bond donor/acceptor motif in practice.The side chain `O[C@H]2CCN(Cc3ccccc3)C2` forms a chiral, constrained, tertiary amine–containing ring bearing a benzyl substituent. This kind of basic nitrogen can be protonated under physiological conditions, enabling salt formation and improving solubility, while the benzyl group adds lipophilicity to support membrane permeation. Overall, the molecule appears to be a medium-sized, heteroatom-rich small molecule with a balance of polar and nonpolar regions typical of orally administered drugs.
Step 2: Qualitative assessment of "drug-likeness"
From a high-level medicinal chemistry perspective, the dihydropyridine core plus ester and aryl substituents is a well-known scaffold in many orally used agents. The overall size seems to be in the usual small-molecule range (roughly a few hundred Daltons), and the combination of:
- one protonatable amine in the side chain,
- multiple carbonyl and "oxime-like" N–O functionalities,
- and at least one aromatic ring,
suggests a reasonable compromise between aqueous solubility (thanks to the basic nitrogen and polar groups) and membrane permeability (thanks to the aromatic and alkyl portions). This fits the typical profile of drug-like molecules that are optimized for oral administration.
Step 3: Use few-shot examples and scaffold prior
Among the provided examples, the first molecule
`COC(=O)C1=C(C)NC(C)=C(C(=O)OCCN(C)Cc2ccccc2)C1c1cccc([N+](=O)[O-])c1`
shares the same dihydropyridine–aryl framework and an analogous "oxime-like" `[N+](=O)[O-]` aromatic substituent, coupled to a benzylated tertiary amine side chain. It is labeled as <answer>True</answer>, indicating good oral bioavailability. The main differences between that example and the target lie in the exact shape and rigidity of the side chain, but the core pharmacophore and the "oxime-like" aromatic pattern are preserved.
The side chain in the target molecule (`O[C@H]2CCN(Cc3ccccc3)C2`) even introduces a chiral, cyclic constraint around the amine, a design element often seen in modern oral drugs to improve receptor binding and ADME properties. Overall, the target structure looks very close to **known dihydropyridine-based, Dapoxetine-like oral agents**: an electron-rich aromatic ring with an "oxime-type" N–O substituent, a dihydropyridine core, and a benzylated amino side chain. **Such scaffolds are widely reported and optimized as orally bioavailable drugs.**
Step 4: Conclusion
Based on its oxime-like aromatic substituent, drug-like dihydropyridine scaffold, and resemblance to known [illegible]ive Dapoxetine-like molecules, I predict that this compound will show oral bioavailability.</think>
<answer>True</answer>

Figure 8: Failed case on MPPREASONER (Text Only) for oral bioavailability prediction (OOD). Same input as Figure 7.