RSI-MOL: RECURSIVE SELF-IMPROVEMENT FOR MOLECULAR PROPERTY PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although molecular property prediction has advanced substantially through large-scale pretraining and sophisticated neural architectures, most existing approaches remain static, with limited ability to learn from prediction failures and refine the task-relevant knowledge and representations that guide subsequent predictions. We argue that molecular property prediction can instead be formulated as recursive self-improvement, where execution feedback is used not only prediction outcomes but also the molecular reasoning state used to generate subsequent predictions. To this end, we introduce **RSI-Mol**, a recursive self-improving framework for molecular property prediction that iteratively evolves its molecular reasoning state through interactions between large language models (LLMs) and downstream predictive models. RSI-Mol realizes recursive self-improvement through two coupled processes: in a self-improvement proposal process, LLMs reason over task-specific knowledge and the current molecular reasoning state to propose task-relevant physicochemical features and predictive substructures for downstream prediction; in a recursive feedback update process, prediction errors and SHAP-based attribution signals produced by the downstream predictor are fed back into the reasoning process to assess the current proposals and guide the refinement, elimination, and discovery of molecular features and substructures. The resulting updated molecular reasoning state, including task features, predictive substructures, feature pools, and failure-case context, is then reused as the starting point of the next reasoning round, where a new downstream predictor is constructed and evaluated. In this way, execution outcomes from one iteration directly update the reasoning state that governs the next, forming a recursive loop of proposal generation, prediction, diagnosis, and state refinement. Beyond improving predictive performance, RSI-Mol provides interpretable molecular evidence through attribution-guided feedback and chemically meaningful substructure refinement. Extensive experiments on eight MoleculeNet benchmarks demonstrate that RSI-Mol achieves the highest average AUC of 81.21% among the compared methods across single-task and multi-task settings. Code and implementation details are available at `https://anonymous.4open.science/r/RSI-Mol-E767/`.

1 INTRODUCTION

Molecular property prediction is typically formulated as a static learning problem: a predictive system is optimized or configured before deployment and then applied to unseen molecules with a largely fixed model, representation, and inference procedure. Recent molecular reasoning systems have substantially improved this paradigm through biomolecular instruction tuning (Fang et al., 2024), few-shot molecular property prediction (Liu et al., 2024), and explicit structural reasoning (Jang et al., 2025). However, these advances primarily improve the model, representation, or reasoning procedure used to produce predictions, while the resulting inference process remains largely fixed once deployed. In their reported formulations, the outcome of an individual prediction does not itself produce a persistent update to the system state that governs subsequent predictions.

Recursive self-improvement (RSI) describes a class of systems that convert experience and evaluation feedback into persistent changes that are inherited by subsequent system states, enabling the system to improve across successive cycles rather than treating execution independently (Liu et al.,

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

2026). More importantly, recursion arises when the updated state not only improves subsequent performance but also influences how future improvements are generated, forming a continuing loop between execution, feedback, and system modification (Duan et al., 2026; Liu et al., 2026). Recent systems such as Dream-RSI and MetaRSI further demonstrate how this principle can be instantiated through persistent modification across repeated improvement cycles (Zheng et al., 2026; Tan et al., 2026). Molecular property prediction provides a particularly grounded setting for studying this form of self-improvement: prediction outcomes can be directly evaluated against labeled properties, while molecular knowledge can be expressed through interpretable descriptors and structural patterns. These properties provide both verifiable feedback and a concrete state in which useful experience can be accumulated and inherited across successive improvement cycles.

Yet, the ingredients that make molecular property prediction well suited to self-improvement are not fully exploited by existing molecular reasoning systems. Iterative LLM-based approaches can progressively refine scientific hypotheses through retrieval, self-critique, or multi-aspect feedback (Wang et al., 2024; Madaan et al., 2023; Yang et al., 2024b; 2025), but these refinements are primarily performed in the language space. Consequently, improved coherence or plausibility of a hypothesis does not necessarily imply that it captures molecular patterns that are actually predictive for the task, leaving the reasoning process vulnerable to plausible but weakly data-supported conclusions. A complementary line of work grounds LLM reasoning with structured domain knowledge, including causal graphs, knowledge graphs, and explicit molecular structural information (Tong et al., 2024; Xiong et al., 2024; Jang et al., 2025). Such grounding improves the connection between language-based reasoning and molecular evidence, but the incorporated knowledge is typically predefined, retrieved, or extracted independently of the specific errors made by the predictive system. It therefore does not directly resolve which aspects of the current molecular representation are responsible for task-specific failures or how that representation should be revised in response.

To address these limitations, we introduce **RSI-Mol**, a recursive self-improving framework for molecular property prediction built around a persistent molecular reasoning state. The state maintains the task-specific features, predictive substructures, feature pool, and accumulated failure context that support downstream prediction. In each improvement cycle, the current state is used to construct a predictive model, whose outputs are evaluated to obtain execution feedback from prediction errors, performance signals, and feature attribution. This evidence is then used to revise the molecular reasoning state by retaining useful knowledge, removing ineffective components, and incorporating newly identified task-relevant information. The resulting successor state becomes the starting point for constructing the predictive model in the next cycle. Through this repeated transition from reasoning state to model execution, diagnostic feedback, and state revision, RSI-Mol allows knowledge acquired in one cycle to directly shape both prediction and model construction in subsequent cycles.

The main contributions of this work are summarized as follows:

1. We formulate a recursive self-improvement (RSI) paradigm for molecular reasoning that continuously transforms execution feedback into persistent updates of molecular knowledge and reasoning strategies, enabling each improvement cycle to shape how subsequent improvements are generated.
2. We propose RSI-Mol, an interpretable framework for molecular property prediction that makes the evolution of molecular hypotheses, structural evidence, and reasoning decisions explicit throughout the self-improvement process, enabling transparent and chemically grounded prediction.
3. Extensive experiments across multiple molecular property benchmarks demonstrate that RSI-Mol consistently improves predictive performance while providing explicit and interpretable structural rationales for its predictions.

2 Related Work

Recursive Self-Improvement Recursive self-improvement (RSI) broadly refers to a process in which a system uses information obtained from its own execution to improve the state that governs subsequent behavior. In this way, the system progressively adapts its internal knowledge, representations, or decision strategies through repeated interaction with downstream feedback. RSI-Mol

brings this idea into molecular property prediction by allowing downstream predictive feedback to directly reshape the molecular features and substructures used in subsequent reasoning rounds.

LLMs for Molecular Reasoning Large language models such as Gemini (Team et al., 2023),GPT-4 (Achiam et al., 2023), Claude, Galactica (Taylor et al., 2022), and BioMedLM (Bolton et al., 2024) have demonstrated exceptional capability in interpreting, synthesizing, and generating scientific content. Consequently, these models have been adopted across multiple stages of the research workflow, including hypothesis discovery, experimental design, scientific writing, and peer review (Luo et al., 2025).

Recent efforts have leveraged LLMs for hypothesis generation, which can be broadly categorized into two groups: *literature-based synthesis* (Yang et al., 2024b; 2025) and *inductive reasoning* (Yang et al., 2024a; Majumder et al., 2024; 2025). Literature-based approaches utilize LLMs' proficiency in fusing cross-document knowledge to uncover latent connections across corpora (Yang et al., 2024b; Tshitoyan et al., 2019). However, such methods primarily rely on textual co-occurrence patterns and often lack task-specific grounding or systematic validation against empirical evidence. In contrast, inductive reasoning methods derive abstracted rules from structured or observational datasets(Qiu et al., 2023; Wang et al., 2023; Zhong et al., 2023; Zhu et al., 2023), offering stronger data grounding but typically remaining sensitive to dataset biases and limited in their ability to generalize across tasks or adapt to evolving domain knowledge. To overcome these limitations, hybrid paradigms have emerged that combine elements of both approaches. For example, MOOSE-Chem (Yang et al., 2025) combines multi-agent workflows with both textual retrieval and structural reasoning to rediscover chemistry hypotheses, often aligning with expert-derived mechanisms. Despite these advances, existing systems rarely close the recursive loop between molecular reasoning and downstream execution: generated knowledge guides prediction, but prediction outcomes do not continuously reshape the state used for subsequent molecular reasoning. RSI-Mol addresses this gap by feeding downstream predictive feedback back into the reasoning process, enabling task-relevant molecular knowledge to evolve across successive rounds.

Molecular Property Prediction Molecular property prediction has evolved from handcrafted features to modern representation learning frameworks(Jiang et al., 2025). Early models relied on extended connectivity fingerprints with classifiers such as random forests(Rogers & Hahn, 2010; Xiong et al., 2019). The introduction of graph neural networks (GNNs) enabled structural learning from molecular graphs, as exemplified by GraphCL (You et al., 2020), GraphMVP (Liu et al., 2021), GEM (Fang et al., 2022), and UniMol (Zhou et al., 2023). These models often use contrastive learning or spatial encoding to enhance generalization. To further enrich molecular representations, multimodal frameworks integrate both structure and text. Recent models include KV-PLM (Zeng et al., 2022), MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2023b), MolCA (Liu et al., 2023c), GIT-Mol (Liu et al., 2023a), BioT5 (Pei et al., 2023), and MOV-MOL (Luo et al., 2024), many of which leverage large-scale datasets of SMILES sequences, scientific abstracts, or ontologies. LLM4SD (Zheng et al., 2025) exemplifies the trend of embedding LLM-derived rules into molecular prediction tasks, imitating expert analytical strategies. RSI-Mol differs from conventional molecular predictors by treating molecular representation construction as an evolving process. Instead of producing a fixed feature representation once, RSI-Mol repeatedly revises task features and predictive substructures according to downstream execution feedback, so that each reasoning round starts from the state accumulated in previous rounds.

3 METHOD

RSI-Mol maintains a persistent *molecular reasoning state* that is updated after each round of downstream prediction. The state records the active molecular representation, the hypotheses that have already been rejected, and the molecular cases used for subsequent structural reasoning. At each round, the current state conditions new task-feature and substructure proposals; these are evaluated by a downstream predictor; and prediction errors together with attribution feedback are written back into the state, which initializes the next round.

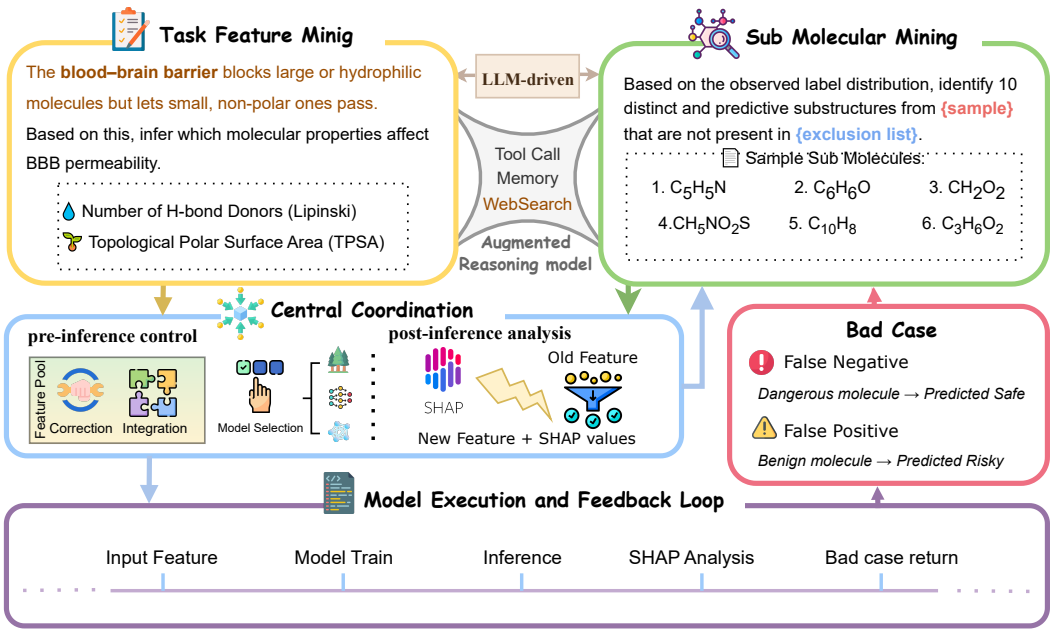


Figure 1: RSI-Mol recursively refines a molecular representation through a persistent reasoning state. Task knowledge is used to derive task-aware physicochemical features, while predictive substructures are proposed from the current molecular context under constraints imposed by previously rejected candidates. The resulting representation is evaluated by a downstream predictor, whose feedback updates the state in two ways: SHAP attribution prunes weak features and expands the exclusion archive, while poorly predicted molecules are retained as failure-case context for the next round. These updated quantities are inherited by subsequent rounds, forming a proposal–execution–feedback loop that progressively changes both the active representation and the conditions under which new structural hypotheses are generated.

3.1 The Molecular Reasoning State

At round r, RSI-Mol maintains

$$\Sigma^{(r)} = \langle \mathcal{T},\ \mathcal{P}^{(r)},\ \mathcal{E}^{(r)},\ \mathcal{D}^{(r)} \rangle, \tag{1}$$

where $\mathcal{T}$ is a task mechanism summary describing the biochemical or pharmacological basis of the target property; $\mathcal{P}^{(r)} = \mathcal{F}^{(r)} \cup \mathcal{S}^{(r)}$ is the active feature pool, partitioned into task-aware physicochemical features $\mathcal{F}^{(r)}$ and predictive substructures $\mathcal{S}^{(r)}$; $\mathcal{E}^{(r)}$ is the exclusion archive, the monotonically growing set of candidates rejected at screening or pruned after execution; and $\mathcal{D}^{(r)}$ is the molecular context on which structural reasoning is currently conditioned. Task-aware features are computed with cheminformatics tools, while substructure features are SMARTS-based match counts. Each round consists of proposal (Sec. 3.2), execution and feedback update (Sec. 3.3). The updated state initializes the next round.

3.2 Self-Improvement Proposal

The proposal process uses an LLM π_θ as a conditional generator of candidate features, and cheminformatics tools as a deterministic admissibility filter. The division of labor is strict: π_θ supplies chemical hypotheses, and nothing it proposes enters the representation without passing a tool-verifiable check.

Task grounding and feature derivation. At the initial round, RSI-Mol retrieves task-relevant information from scientific and biomedical sources and uses the LLM π_θ to summarize the underlying biological or physicochemical mechanism into a task description $\mathcal{T}$. Conditioned on $\mathcal{T}$, π_θ then derives a set of task-relevant molecular descriptors $\mathcal{F} = \{f_1, \ldots, f_m\}$, which are instantiated using

cheminformatics tools and subsequently refined by downstream attribution feedback. Examples of task reasoning and feature derivation are provided in Appendix F.

State-conditioned substructure proposal. Substructure generation is performed at every round and conditioned on the current reasoning state. At round r, candidates are generated sequentially as

$$s_j^{(r)} \sim \pi_\theta\Big(\cdot \mid \mathcal{T}, \mathcal{D}^{(r)}, \mathcal{E}^{(r)}, \mathcal{S}^{(r)} \cup \{s_1^{(r)}, \ldots, s_{j-1}^{(r)}\}\Big), \tag{2}$$

where $\mathcal{D}^{(0)}$ is initialized from sampled labeled molecules and, in later rounds, is replaced by molecules that the current predictor fails to model well. The exclusion set $\mathcal{E}^{(r)}$ prevents previously rejected or pruned candidates from being proposed again, while the partial substructure set avoids within-round duplication. Generated SMARTS patterns are validated with RDKit before being added to the current feature pool; invalid or redundant candidates are discarded. Prompt templates and screening details are provided in Appendix F.

3.3 Recursive Feedback Update

At round r, the predictor $M^{(r)}$ is trained on $\mathcal{Z}_{\mathrm{tr}}$ using the current feature pool and evaluated on $\mathcal{Z}_{\mathrm{val}}$. Each execution produces two feedback signals that update different parts of the molecular reasoning state.

Attribution-guided representation update. SHAP-based attribution is used to identify weakly contributing features, which are removed from the active representation:

$$\mathcal{P}^{(r+1)} = \widehat{\mathcal{P}}^{(r)} \setminus \mathcal{R}^{(r)}, \tag{3}$$

where $\mathcal{R}^{(r)}$ denotes the features selected for pruning according to the attribution criterion. Pruned features, together with candidates rejected during structural screening, are added to the exclusion archive,

$$\mathcal{E}^{(r+1)} = \mathcal{E}^{(r)} \cup \mathcal{R}^{(r)} \cup \mathcal{R}_{\mathrm{scr}}^{(r)}. \tag{4}$$

Because ($\mathcal{E}^{(r+1)}$) is used to condition subsequent substructure proposals, attribution feedback updates both the retained representation and the structural hypotheses considered in later rounds. Detailed SHAP definitions, attribution statistics, and pruning thresholds are provided in Appendix B.

Error-guided context update. In parallel, RSI-Mol collects a failure set $\mathcal{B}^{(r)}$ containing poorly predicted validation molecules, including false positives and false negatives for classification and large-residual examples for regression. These cases are not used for feature pruning. Instead, the top-k failure cases are selected as the molecular context for the next round,

$$\mathcal{D}^{(r+1)} = \mathrm{TopK}\big(\mathcal{B}^{(r)}, k\big), \qquad k = 5, \tag{5}$$

and are supplied to π_θ during subsequent substructure generation. Prediction errors therefore directly change the molecular evidence used to generate the next set of structural hypotheses.

The refinement process is repeated for at most R rounds and terminates early when no new candidate is admitted and no existing feature is removed.

3.4 The Recursive Structure of RSI-Mol

The recursive aspect of RSI-Mol arises from three inherited quantities, which differ in kind. The retained feature pool $\mathcal{P}^{(r+1)}$ determines what the next predictor is able to see (*representational* inheritance, Eq. equation 3). The failure cases $\mathcal{D}^{(r+1)}$ determine which molecules the next round of structural reasoning must explain (*evidential* inheritance, Eq. equation 5). The exclusion archive $\mathcal{E}^{(r+1)}$ constrains what may be proposed at all (*procedural* inheritance, Eq. equation 4). The third is the one that makes the loop recursive rather than merely repeated: because $\mathcal{E}$ enters the conditioning set of Eq. equation 2, execution feedback modifies not only the current representation but the conditions under which the next representation is generated.

4 EXPERIMENTS

Datasets We evaluate RSI-Mol on eight benchmark datasets from MoleculeNet (Wu et al., 2018), covering a diverse range of molecular prediction tasks. Following prior work (Luo et al., 2024), we split each dataset into training, validation, and test sets (8:1:1) using a scaffold-based strategy to ensure structural diversity. For multi-task datasets with severe label imbalance, tasks whose test sets exhibit label degeneracy (i.e., containing only a single class) are excluded.

Baselines We compare the proposed RSI-Mol against multiple baselines, encompassing both unimodal and multimodal molecular modeling methods, including unimodal baselines: GraphCL (You et al., 2020), GraphMVP (Liu et al., 2021), GEM (Fang et al., 2022), UniMol (Zhou et al., 2023), and multimodal baselines: KV-PLM (Zeng et al., 2022), MoMu (Su et al., 2022), MoleculeSTM (Liu et al., 2023b), MolCA (Liu et al., 2023c), GIT-Mol (Liu et al., 2023a), BioT5 (Pei et al., 2023), MV-Mol (Luo et al., 2024), SCAGE (Qiao et al., 2025), MOlCL-SP (Luo & Deng, 2025).

Implementation Details To comprehensively evaluate the performance of our model on molecular property prediction tasks, we adopt the widely used evaluation metric: area under the curve (AUC) across all eight datasets. All large model inferences are executed using DeepSeek-R1 (Guo et al., 2025). In each reasoning iteration, we select 10 candidate molecular substructures and 5 task-specific molecular features. Error correction is managed via RDKit-based molecular validity checks to ensure chemical plausibility. Model selection is automated via AutoGluon (Erickson et al., 2020), using cross-validation performance to identify optimal configurations. The refinement process is run for at most 30 iterations, with feature elimination performed periodically based on feature activity. At each iteration, 5 error cases are returned for analysis. All framework operations are conducted exclusively on the training and validation sets, and the test set is used solely for final performance evaluation.

4.1 OVERALL MOLECULAR PROPERTY PREDICTION PERFORMANCE

Table 1: Molecular property prediction performance comparison across datasets (values: ROC-AUC(%) $\pm$ std). The best performances are marked in bold. ↑: higher scores are better. -: The score was not reported in the original publication.

Model	BBBP ↑	Tox21 ↑	ToxCast ↑	SIDER ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	Avg ↑
GraphCL	$67.5_{\pm 3.3}$	$75.0_{\pm 0.3}$	$62.8_{\pm 0.2}$	$60.1_{\pm 1.3}$	$78.9_{\pm 4.2}$	$77.1_{\pm 1.0}$	$75.0_{\pm 0.4}$	$68.7_{\pm 7.8}$	70.64
GraphMVP	$72.4_{\pm 1.6}$	$74.4_{\pm 0.2}$	$63.1_{\pm 0.4}$	$63.9_{\pm 1.2}$	$77.5_{\pm 4.2}$	$75.0_{\pm 1.0}$	$77.0_{\pm 1.2}$	$81.2_{\pm 0.9}$	73.07
GEM	$72.4_{\pm 0.4}$	$78.1_{\pm 0.1}$	$69.2_{\pm 0.4}$	$67.2_{\pm 0.4}$	$90.1_{\pm 1.3}$	$81.7_{\pm 0.5}$	$80.6_{\pm 0.9}$	$85.6_{\pm 1.1}$	78.11
Uni-Mol	$72.9_{\pm 0.6}$	$79.6_{\pm 0.5}$	$69.6_{\pm 0.1}$	$65.9_{\pm 1.3}$	$91.9_{\pm 1.8}$	$82.1_{\pm 1.3}$	$80.8_{\pm 0.3}$	$85.7_{\pm 0.2}$	78.56
SCAGE	$73.4_{\pm 1.1}$	$79.4_{\pm 1.2}$	$69.3_{\pm 0.5}$	$66.0_{\pm 1.2}$	$92.7_{\pm 0.9}$	-	-	$85.4_{\pm 1.2}$	-
KV-PLM	$66.9_{\pm 1.1}$	$64.7_{\pm 1.8}$	$58.6_{\pm 0.4}$	$55.3_{\pm 0.8}$	$84.3_{\pm 1.5}$	$60.2_{\pm 2.9}$	$68.8_{\pm 4.9}$	$71.9_{\pm 2.1}$	66.29
MoMu	$70.5_{\pm 2.0}$	$75.6_{\pm 0.3}$	$63.4_{\pm 0.5}$	$60.5_{\pm 0.9}$	$79.9_{\pm 4.1}$	$70.5_{\pm 1.4}$	$75.9_{\pm 0.8}$	$76.7_{\pm 2.1}$	71.63
MoleculeSTM	$70.0_{\pm 0.5}$	$77.0_{\pm 0.4}$	$65.1_{\pm 0.3}$	$61.0_{\pm 1.0}$	$92.5_{\pm 1.1}$	$73.4_{\pm 2.9}$	$76.9_{\pm 1.8}$	$80.8_{\pm 1.3}$	74.57
MolCA	$70.0_{\pm 0.5}$	$77.2_{\pm 0.5}$	$64.5_{\pm 0.8}$	$63.0_{\pm 1.7}$	$89.5_{\pm 0.7}$	-	-	$79.8_{\pm 0.5}$	-
GIT-Mol	$73.9_{\pm 0.6}$	$75.9_{\pm 0.5}$	$66.8_{\pm 0.5}$	$63.4_{\pm 0.8}$	$88.3_{\pm 1.2}$	-	-	$81.1_{\pm 1.5}$	-
BioT5	$75.2_{\pm 0.1}$	$75.5_{\pm 0.3}$	$64.4_{\pm 0.1}$	$62.4_{\pm 0.1}$	$85.8_{\pm 0.6}$	$75.9_{\pm 0.7}$	$79.7_{\pm 0.4}$	$\mathbf{89.2}_{\pm 0.1}$	76.01
MolCL-SP	$74.4_{\pm 0.7}$	$76.9_{\pm 1.0}$	$68.4_{\pm 0.2}$	$65.6_{\pm 0.1}$	$\mathbf{99.4}_{\pm 0.2}$	$82.2_{\pm 0.8}$	$80.9_{\pm 0.5}$	$82.9_{\pm 1.0}$	78.84
MV-Mol	$73.6_{\pm 0.2}$	$80.3_{\pm 0.6}$	$70.0_{\pm 0.4}$	$67.3_{\pm 0.0}$	$95.6_{\pm 1.6}$	$82.1_{\pm 0.5}$	$81.4_{\pm 0.3}$	$88.2_{\pm 0.4}$	79.80
RSI-Mol	$\mathbf{77.1}_{\pm 0.2}$	$\mathbf{81.0}_{\pm 0.4}$	$\mathbf{71.2}_{\pm 0.5}$	$\mathbf{69.1}_{\pm 0.1}$	$95.8_{\pm 0.7}$	$\mathbf{84.3}_{\pm 0.4}$	$\mathbf{82.3}_{\pm 0.3}$	$88.9_{\pm 0.2}$	**81.21**

Table 1 summarizes the performance of RSI-Mol compared to existing baseline models across eight benchmark datasets.

RSI-Mol consistently outperforms all competitors, achieving the highest overall average AUC score of 81.21%. It exceeds the best-performing unimodal model UniMol (78.56%) by 2.65% and surpasses the strongest multimodal model MV-Mol (79.80%) by 1.41%. In the single-task setting, RSI-Mol delivers notable improvements. On the BBBP dataset, it achieves an AUC of 77.1%, outperforming BioT5 (75.2%), which was previously the strongest model on this task. On ClinTox, RSI-Mol achieves an AUC of 95.8%, slightly higher than MV-Mol (95.6%) and substantially outperforming UniMol (91.9%), demonstrating robust performance in binary toxicity prediction tasks

relevant to pharmacological safety assessment. In the multi-task setting, RSI-Mol maintains strong generalization capabilities. On the MUV dataset, it achieves an AUC of 84.3%, outperforming MV-Mol by 2.2%. On SIDER, RSI-Mol reaches 69.1%, compared to MV-Mol's 67.3%. RSI-Mol also achieves the best result on Tox21 with an AUC of 81.0%, improving upon the previous best of 80.3% by MV-Mol. These findings indicate that RSI-Mol effectively scales to multi-task prediction challenges and benefits from its task-guided iterative reasoning mechanism.

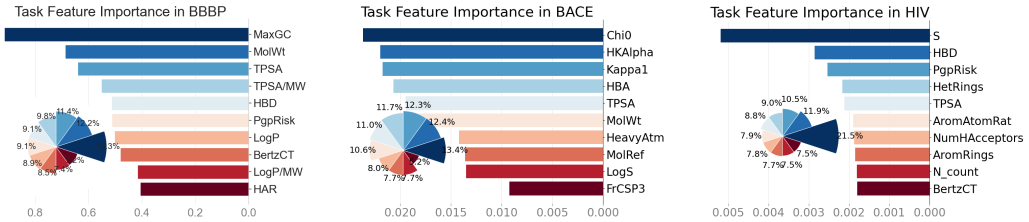


Figure 2: Task-aware feature importance after recursive refinement on the BBBP, BACE, and HIV datasets. The x-axis represents the mean absolute SHAP value, and the y-axis lists the 10 most influential retained features. The accompanying pie charts show their relative contributions.

Task-Aware Features Retained by Recursive Refinement Figure 2 shows the task-aware features retained after recursive refinement for BBBP, BACE, and HIV, ranked by their mean absolute SHAP values. The resulting feature profiles exhibit clear task-specific patterns, indicating that RSI-Mol adaptively preserves different molecular descriptors according to the downstream prediction objective. For BBBP, the most influential features are MaxGC, TPSA, and MolWt[1], which are closely related to blood–brain barrier permeability and capture key physicochemical properties affecting molecular transport (Gupta et al., 2019). For BACE, Chi0, HallKierAlpha, and Kappa1 contribute most strongly, reflecting molecular shape and flexibility relevant to interactions with the BACE1 binding site (Guo et al., 2020). For HIV, sulfur presence has the largest influence on model predictions, consistent with reported associations between sulfur-containing groups and antiviral activity (Honarmand Ebrahimi et al., 2022; Grimble, 2006; Akor et al., 2024). Importantly, the dominant features differ substantially across the three tasks rather than converging to a common generic descriptor set. This suggests that the recursive refinement process is task-sensitive: feedback from downstream prediction reshapes the active representation while retaining features that remain informative for the specific molecular property being modeled. Overall, these results show that RSI-Mol preserves task-relevant molecular descriptors while progressively removing weakly contributing features from the representation.

Rollout dynamics on ClinTox

We conduct 20-step rollouts on ClinTox with three random seeds and record the validation AUC at each step. The solid line shows the mean performance across runs, the shaded region denotes ± 1 standard deviation, and the dashed line indicates the mean AUC at the first step. The validation AUC increases from 74.2% to 88.5% over the rollout, demonstrating that RSI-Mol continues to improve as its reasoning state is recursively updated. Rather than treating each refinement step independently, RSI-Mol carries forward the retained feature representation, failure-case context, and exclusion history from previous rounds, allowing later proposals to build on earlier feedback. The resulting upward trend shows that these state updates accumulate across rounds and progressively strengthen the downstream predictor. This provides direct evidence for the self-improving behavior of RSI-Mol: prediction outcomes from one round are fed back into the reasoning process and used to guide subsequent refinement.

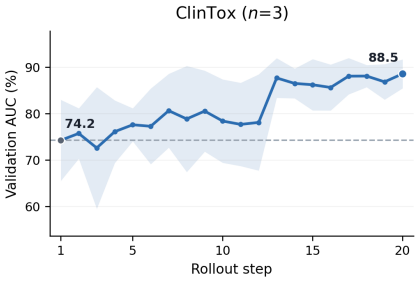


Figure 3: Rollout dynamics on ClinTox.

[1]All abbreviations appearing in the article are explained in the Appendix.

Transferability of Refined Structural Hypotheses To examine whether the structural hypotheses refined by RSI-Mol can transfer across related prediction tasks, we decompose the SIDER dataset into 20 base tasks and 7 derived tasks. Substructures are first obtained from the base tasks through the RSI-Mol refinement process. For each derived task, we reuse these refined substructures and select the subset that provides the strongest predictive performance. As shown in Figure 4, RSI-Mol maintains robust performance across the 27 adverse-effect categories, with particularly strong results on Gastrointestinal, Skin, and Investigation-related tasks. Performance is comparatively lower on categories such as Surgical, whose outcomes may depend more strongly on contextual factors beyond molecular structure. For the 7 derived tasks, the selected substructures do not simply originate from the base task with the highest individual predictive performance. Instead, several transferred patterns are consistent with related biological endpoints. For example, the Neoplasm task selects substructures originating from Breast, while Immune selects features from Skin.

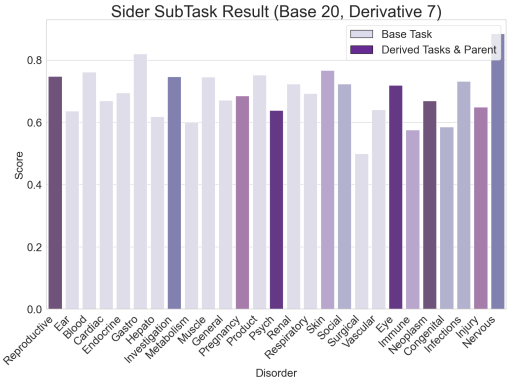


Figure 4: Performance across 27 SIDER adverse effect categories, including 20 base and 7 derived tasks.

4.2 Ablation Study

Ablation of the Active Feature Pool Table 2 evaluates the two components of the active feature pool, task-aware molecular features ($\mathcal{F}$) and predictive substructures ($\mathcal{S}$). Removing both components leads to the lowest performance across all benchmarks, while introducing either component individually improves prediction performance. Using both feature types consistently achieves the strongest results.

Table 2: RSI-Mol feature ablation on HIV, BBBP, ClinTox, and Tox21.

$\mathcal{S}$	$\mathcal{F}$	HIV	BBBP	ClinTox	Tox21
-	-	79.0 (±0.4)	70.5 (±0.2)	84.2 (±1.6)	62.9 (±1.6)
+	-	81.9 (±0.3)	72.1 (±0.1)	87.3 (±0.7)	76.2 (±0.5)
-	+	80.4 (±0.3)	73.2 (±0.2)	87.5 (±0.9)	70.7 (±0.4)
+	+	**82.3 (±0.3)**	**77.1 (±0.2)**	**95.8 (±0.7)**	**81.0 (±0.4)**

For example, on HIV, performance increases from 79.0% without either component to 82.3% when both are included, while BBBP improves from 70.5% to 77.1%. Similar gains are observed on ClinTox and Tox21. These results show that task-aware descriptors and predictive substructures provide complementary information within the refined molecular representation.

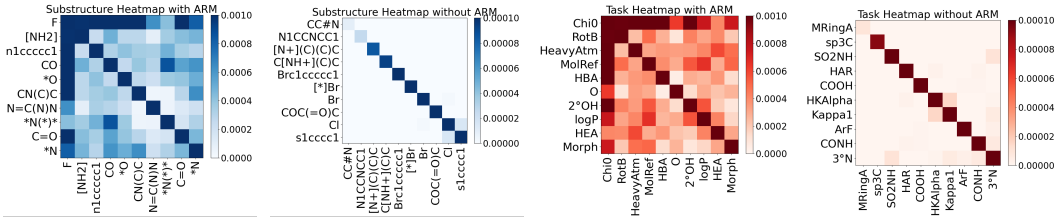


Figure 5: Comparison of feature interaction patterns with and without state-conditioned reasoning on the BACE dataset. The first two panels show substructure-level interactions, while the last two show task-aware feature interactions.

Effect of State-Conditioned Reasoning To evaluate the contribution of the augmented reasoning module—which comprises memory, tool use, and web search components (grey section in Fig-

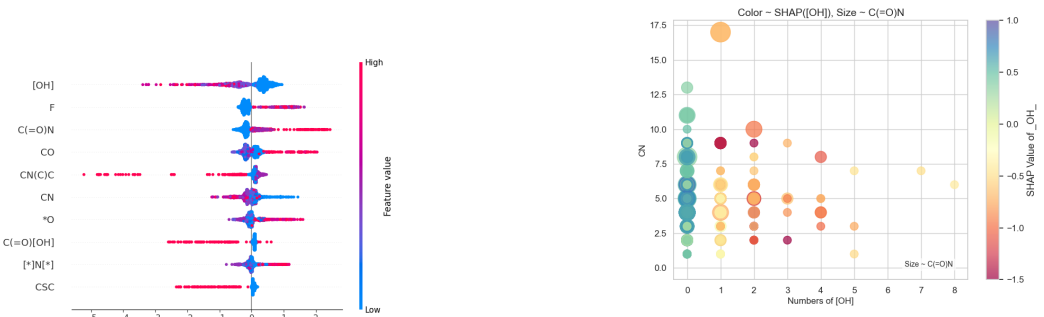
ure 1)—we conducted an ablation study on the BACE dataset by disabling all auxiliary functionalities and retaining only the baseline prompt-based feature extraction. The comparative outcomes are illustrated in Figure 5. The heatmap shows that removing the reasoning module reduces overall feature synergy by an order of magnitude, as measured by SHAP interaction values. This reflects a clear decline in the model's ability to perform coherent and context-aware molecular interpretation, underscoring the role of augmented reasoning in integrating biologically meaningful features for predictive modeling.

Generalization Beyond Molecular Property Prediction To evaluate the generalizability of RSI-Mol beyond molecular property prediction, we extend the framework to drug–target interaction (DTI) prediction on DrugBank. Following the data split of R-DTI, we replace task-aware molecular feature mining with protein fragment mining, while keeping the recursive proposal, execution, and feedback procedure unchanged. As shown in Table 3, RSI-Mol achieves an AUC-ROC of 91.7% and an AUC-PR of 92.8%, outperforming the compared DTI baselines. These results indicate that the recursive reasoning framework can be adapted to a different molecular prediction setting by modifying the task-specific proposal component while preserving the overall refinement process.

Table 3: Performance comparison on the DrugBank dataset.

Method	AUC-ROC	AUC-PR
TriMulDTI (Dehghan et al., 2023)	89.5 (±0.2)	89.7 (±0.3)
MFR-DTA (Hua et al., 2023)	89.3 (±0.2)	89.1 (±0.2)
PSC-CPI (Wu et al., 2024)	89.5 (±0.1)	89.7 (±0.1)
R-DTI (Hua et al., 2025)	91.4 (±0.1)	91.2 (±0.1)
RSI-Mol	**91.7 (±0.1)**	**92.8 (±0.1)**

4.3 How Attribution Feedback Shapes the Representation



(a) SHAP summary of key substructures.

(b) Joint attribution of representative substructures.

Figure 6: SHAP analysis of key substructures for BBB permeability. (a) Distribution of SHAP values for influential substructures. (b) Effects of the occurrence counts of representative positive (C(=O)N), negative ([OH]), and neutral (CN) substructures on prediction.

Figure 6 presents the SHAP analysis of representative substructures on BBBP. [OH], C(=O)[OH], and CSC mainly contribute negatively to BBB permeability, while C(=O)N shows positive effects for some molecules. In particular, the consistently negative contribution of [OH] agrees with the known influence of polarity and hydrogen-bonding on BBB diffusion (Abraham, 2004), whereas C(=O)N and CN exhibit more context-dependent effects. These attribution patterns provide feedback for refining the active representation in RSI-Mol.

5 Conclusion

We presented **RSI-Mol**, a recursive self-improving framework for molecular property prediction. Through iterative updates of the molecular reasoning state, RSI-Mol uses attribution and prediction feedback to progressively refine task-aware features and structural hypotheses. The results show that this recursive refinement improves prediction performance while maintaining interpretable and biologically meaningful representations.

6 AI USE STATEMENT

Generative AI tools were used during the preparation of this manuscript to assist with language polishing, grammatical correction, and improving the clarity and readability of the text. All AI-assisted suggestions were reviewed and revised by the authors.

REFERENCES

Michael H Abraham. The factors that influence permeation across the blood–brain barrier. *European journal of medicinal chemistry*, 39(3):235–240, 2004.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Faith O Akor, Godwin D Edo, Favour A Nelson, Abasifreke U Johnson, Solomon O Iyam, Muhammad N Abubakar, Alpha O Gulack, Chioma B Ubah, Bassey O Ekpong, and Innocent Benjamin. Surface modification of graphene and fullerene with sulfur (s), selenium (se), and oxygen (o): Dft simulation for enhanced zidovudine delivery in hiv treatment. *BMC chemistry*, 18(1):156, 2024.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. Biomedlm: A 2.7b parameter language model trained on biomedical text, 2024. URL `https://arxiv.org/abs/2403.18421`.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Alireza Dehghan, Parvin Razzaghi, Karim Abbasi, and Sajjad Gharaghani. Tripletmultidti: multimodal representation learning in drug-target interaction prediction with triplet loss function. *Expert Systems with Applications*, 232:120754, 2023.

Yi Duan, Ying Liu, Zirui Tang, Haodong Chen, Jun Zhou, Yumou Liu, Bangrui Xu, Yukai Wu, Sidi Chen, Yuhan Zhou, et al. The last ai built by humans: Toward genuine recursive self-improvement. *arXiv preprint arXiv:2609.11873*, 2026.

Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 48221–48251, 2024.

Robert F Grimble. The effects of sulfur amino acid intake on immune function in humans. *The Journal of nutrition*, 136(6):1660S–1665S, 2006.

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

Tiantian Guo, Denghong Zhang, Yuzhe Zeng, Timothy Y Huang, Huaxi Xu, and Yingjun Zhao. Molecular and cellular mechanisms underlying the pathogenesis of alzheimer's disease. *Molecular neurodegeneration*, 15(1):40, 2020.

Mayuri Gupta, Hyeok Jun Lee, Christopher J Barden, and Donald F Weaver. The blood–brain barrier (bbb) score. *Journal of medicinal chemistry*, 62(21):9824–9836, 2019.

Kourosh Honarmand Ebrahimi, Simone Ciofi-Baffoni, Peter-Leon Hagedoorn, Yvain Nicolet, Nick E Le Brun, Wilfred R Hagen, and Fraser A Armstrong. Iron–sulfur clusters as inhibitors and catalysts of viral replication. *Nature Chemistry*, 14(3):253–266, 2022.

Yang Hua, Xiaoning Song, Zhenhua Feng, and Xiaojun Wu. Mfr-dta: a multi-functional and robust model for predicting drug–target binding affinity and region. *Bioinformatics*, 39(2):btad056, 2023.

Yang Hua, Tianyang Xu, Xiaoning Song, Zhenhua Feng, Rui Wang, Wenjie Zhang, and Xiaojun Wu. R-dti: Drug target interaction prediction based on second-order relevance exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17368–17376, 2025.

Yunhui Jang, Jaehyung Kim, and Sungsoo Ahn. Structural reasoning improves molecular understanding of llm. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21016–21036, 2025.

Pengcheng Jiang, Cao Xiao, Tianfan Fu, Parminder Bhatia, Taha Kass-Hout, Jimeng Sun, and Jiawei Han. Bi-level contrastive learning for knowledge-enhanced molecule representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 352–360, 2025.

Pengfei Liu, Yiming Ren, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *arXiv preprint arXiv:2308.06911*, 2023a.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2021.

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

Shuaiqi Liu, Zhengkai Lin, Yuxiang Zhang, Yuanyi Ren, Yue Wu, Yongbin Li, Zheng Wang, Zhihang Fu, and Jieping Ye. The path to recursive self-improving agents: Foundation, framework, and future directions. 2026.

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

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. Molca: Molecular graph-language modeling with cross-modal projector and unimodal adapter. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023c.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, Zikun Nie, Hao Zhou, and Zaiqing Nie. Learning multi-view molecular representations with structured and unstructured knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2082–2093, 2024.

Yue Luo and Lei Deng. Molcl-sp: a multimodal contrastive learning framework with non-overlapping substructure perturbations for molecular property prediction. *Bioinformatics*, 41(10): btaf507, 2025.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research, 2025. URL `https://arxiv.org/abs/2501.04306`.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. Position: Data-driven discovery with large generative models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=5SpjhZNXtt.

Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vyflgpwfJW.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1123, 2023.

Jianbo Qiao, Junru Jin, Ding Wang, Saisai Teng, Junyu Zhang, Xuetong Yang, Yuhang Liu, Yu Wang, Lizhen Cui, Quan Zou, et al. A self-conformation-aware pre-training framework for molecular property prediction with substructure interpretability. *Nature Communications*, 16(1): 4382, 2025.

Linlu Qiu, Liwei Jiang, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, et al. Phenomenal yet puzzling: Testing inductive reasoning capabilities of language models with hypothesis refinement. *arXiv preprint arXiv:2310.08559*, 2023.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481*, 2022.

Zihan Tan, Leixin Sun, Zitong Shi, Yitao Liu, Jiajun Wu, Nathaniel Brooks, Jiaru Qian, Xiaoran Shang, Suyuan Huang, Yi Ding, et al. Metarsi/rsi2: A meta-recursive self-improving system for recursive self-improving systems themselves. *arXiv preprint arXiv:2609.06396*, 2026.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science, 2022. URL https://arxiv.org/abs/2211.09085.

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

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating psychological hypothesis generation with AI: when large language models meet causal graph. *Humanities and Social Sciences Communications*, 11(1):896, July 2024.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

Ruocheng Wang, Eric Zelikman, Gabriel Poesia, Yewen Pu, Nick Haber, and Noah D Goodman. Hypothesis search: Inductive reasoning with language models. *arXiv preprint arXiv:2309.05660*, 2023.

Lirong Wu, Yufei Huang, Cheng Tan, Zhangyang Gao, Bozhen Hu, Haitao Lin, Zicheng Liu, and Stan Z. Li. Psc-cpi: Multi-scale protein sequence-structure contrasting for efficient and generalizable compound-protein interaction prediction, 2024. URL https://arxiv.org/abs/2402.08198.

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

Guangzhi Xiong, Eric Xie, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*, 2024.

Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xiaozhe Wan, Xutong Li, Zhaojun Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *Journal of medicinal chemistry*, 63(16):8749–8760, 2019.

Zonglin Yang, Li Dong, Xinya Du, Hao Cheng, Erik Cambria, Xiaodong Liu, Jianfeng Gao, and Furu Wei. Language models as inductive reasoners. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209–225, St. Julian's, Malta, March 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.13. URL https://aclanthology.org/2024.eacl-long.13/.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13545–13565, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.804. URL https://aclanthology.org/2024.findings-acl.804/.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=X9OfMNNepI.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(1):862, 2022.

Tong Zheng, Xidong Wu, Zheng Zhang, Zhankui He, Chaoyi Zhang, Benjamin Coleman, Ruoqiao Wei, Di Bai, Haolin Liu, Rui Liu, et al. Dream-rsi: Recursive self-improvement through evolving worlds. *arXiv preprint arXiv:2609.14858*, 2026.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T N Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, 7(3):437–447, March 2025.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237, 2023.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: a universal 3d molecular representation learning framework. 2023.

Zhaocheng Zhu, Yuan Xue, Xinyun Chen, Denny Zhou, Jian Tang, Dale Schuurmans, and Hanjun Dai. Large language models can learn rules. *arXiv preprint arXiv:2310.07064*, 2023.

A AutoGluon Ablation

To quantify the contribution of LLM-driven knowledge integration, we compared **RSI-Mol** against AutoGluon trained with the full molecular feature table, including all task values and molecular fingerprints, without any LLM-based filtering or reasoning-driven pruning. As summarized in Table 4, RSI-Mol consistently outperforms the baseline across all benchmarks, achieving an average improvement of 9.3% in AUC. This confirms that the iterative orchestration of LLMs for domain knowledge mining is essential for capturing complex chemical logic that raw data alone cannot provide.

Table 4: RSI-Mol vs. AutoML without LLM-based filtering.

Dataset	AutoGluon + Fingerprint	RSI-Mol
BBBP	66.4 (±0.2)	**77.1 (±0.2)**
ClinTox	77.8 (±0.9)	**95.8 (±0.7)**
HIV	79.1 (±0.5)	**82.3 (±0.4)**
BACE	84.5 (±0.3)	**88.9 (±0.2)**
Tox21	72.9 (±0.5)	**81.0 (±0.4)**
MUV	72.2 (±0.2)	**84.3 (±0.4)**
SIDER	61.8 (±0.1)	**69.1 (±0.1)**
ToxCast	60.2 (±0.4)	**71.2 (±0.5)**
AVG	71.9	**81.2**

Generalizability and Task Extensibility To evaluate RSI-Mol's generalizability, we extended it to the Drug-Target Interaction (DTI) task. Following R-DTI's dataset splitting on DrugBank, we maintained a consistent setup for fair comparison. Leveraging the modular nature of RSI-Mol, we made minimal adjustments to the prompting strategy. Specifically, the Task Feature Mining module was replaced with Protein Fragment Mining to capture relevant biological features. The core Central Coordination and Model Execution and Feedback Loop remained architecturally consistent, utilizing machine learning models to perform adaptive reasoning on the newly integrated knowledge. As shown in Table 3, RSI-Mol outperforms the leading R-DTI model with 0.917 AUC-ROC and 0.928 AUC-PR, with the 1.6% improvement in AUC-PR.

B SHAP-Based Attribution and Feature Pruning

At each refinement round, RSI-Mol applies SHAP analysis to quantify the contribution of each active feature in the current molecular representation. Let $\widehat{\mathcal{P}}^{(r)}$ denote the feature pool evaluated at round r, and let $\phi_k(\varphi, \varphi')$ denote the SHAP interaction value between features φ and φ' for validation molecule k.

For each feature $\varphi \in \widehat{\mathcal{P}}^{(r)}$, we compute its average main effect as

$$v_\varphi^{(r)} = \frac{1}{N} \sum_{k=1}^{N} |\phi_k(\varphi, \varphi)| \,, \tag{6}$$

where $N = |\mathcal{Z}_{\text{val}}|$.

We additionally compute its interaction contribution as

$$c_\varphi^{(r)} = \sum_{\varphi' \in \widehat{\mathcal{P}}^{(r)}} \frac{1}{N} \sum_{k=1}^{N} |\phi_k(\varphi, \varphi')| - v_\varphi^{(r)}. \tag{7}$$

The main effect captures the direct contribution of a feature to prediction, whereas the interaction effect captures its contribution through interactions with other features. Considering both terms avoids removing features that have limited individual importance but remain useful through feature interactions.

A feature is marked for pruning when both its main and interaction effects fall below predefined thresholds:

$$\mathcal{R}^{(r)} = \left\{ \varphi \in \widehat{\mathcal{P}}^{(r)} \;\middle|\; v_\varphi^{(r)} < \tau_v \;\wedge\; c_\varphi^{(r)} < \tau_c \right\}. \tag{8}$$

The retained feature pool is then updated as

$$\mathcal{P}^{(r+1)} = \widehat{\mathcal{P}}^{(r)} \setminus \mathcal{R}^{(r)}. \tag{9}$$

Features removed through attribution-based pruning are added to the exclusion archive and are not reconsidered during subsequent substructure proposal rounds. This allows attribution feedback to affect both the representation retained by the predictor and the structural search performed in later iterations.

The thresholds τ_v and τ_c, together with the SHAP implementation details, are reported in the experimental configuration.

C Model Selection Ablation

To evaluate the contribution of dynamic model selection, we perform an ablation study on the HIV and Tox21 datasets. According to the final AutoGluon outputs, K-Nearest Neighbors consistently yields the weakest performance across both tasks, while XGBoost (Chen & Guestrin, 2016) and Random Forest(Breiman, 2001) achieve the best fixed performance on HIV and Tox21, respectively. We compare these static baselines—each initialized with either the best or the worst performing model—to our dynamic per-round model selection strategy. As shown in Table 5, even the strongest static model underperforms the dynamic strategy by approximately 4.6% on both datasets.

Table 5: Performance comparison using fixed versus dynamic model selection. The dynamic approach (AutoGluon) automatically selects and ensembles the best model for each iter.

Type	Model	HIV	Tox21
Fixed	KNN	59.8 (±0.5)	58.6 (±0.8)
	XGBoost	-	76.4 (±0.5)
	Random Forest	77.7 (±0.3)	-
Dynamic	AutoGluon	**82.3 (±0.2)**	**81.0 (±0.4)**

D Robustness to Feature Initialization

Table 6: Impact of initialization strategies on AUC performance.

Method	Initial	After 5 Epochs	After 10 Epochs
K-means	0.70473	0.71820	0.72890
Random	0.70447	0.71726	**0.72895**

To evaluate whether the performance of RSI-Mol depends on the initial feature pool, we conducted a sensitivity analysis comparing two initialization strategies on the BBBP dataset. We compared a **Random** initialization against a **K-means** clustering approach that uses centroids of positive and negative samples as a heuristic starting point.

The results in Table 6 demonstrate that both strategies converge to a nearly identical final performance. This confirms that the framework's efficacy is driven by its iterative selection and mining mechanism rather than the specific distribution of the initial feature pool.

E Cost Analysis

Table 7: Efficiency summary across different LLM backends on the BBBP dataset.

Module	AUC	FM	SM	ME	PI	In ($)	Out ($)
DeepSeek-v3	71.6	19.6s	4.4s	9.6s	7.1s	0.28	0.42
GPT-4o	76.5	17.2s	17.7s	8.2s	7.6s	5.00	15.00
Deepseek-R1	**77.1**	186s	202s	8.1s	7.9s	0.55	2.19

*FM: Feature Mining; SM: Substructure Mining; ME: Model Execution; PI: Post-Inference. Costs are per 1M tokens.

To demonstrate the practical deployment potential of **RSI-Mol**, we evaluate its computational efficiency and scalability across different Large Language Model (LLM) backends. Unlike heavy pre-training frameworks that require massive GPU clusters, RSI-Mol is designed for high efficiency in limited-compute settings. We measured the runtime and token expenditure for each core module using the BBBP dataset. As shown in Table 7, while reasoning-heavy models like DeepSeek-R1 incur higher latency during the mining phases, the overall operational cost remains significantly lower than proprietary models like GPT-4o.

For reasoning-intensive models, we implement a persistent storage strategy for reasoning content. In cases where the reasoning chain exceeds sequence limits, the framework automatically resumes from the previous state, ensuring continuity in complex chemical logic synthesis.

F EXAMPLE ON THE CLINTOX DATASET

TASK FEATURE MINING PROMPT:

Background(from WebSearch):
According to Lipinski's rule, an orally active drug typically has no more than one violation among the following: no more than 5 hydrogen bond donors, 10 hydrogen bond acceptors, a molecular mass under 500 daltons, and a calculated logP not exceeding 5.
Your task is to identify five molecular properties most likely to influence clinical toxicity (CT_TOX) in the Tox21-derived ClinTox dataset.
Constraints:

- Chosen properties must be computable from molecular structure using RDKit.
- Do not include trivial, redundant, or non-specific properties (e.g., "toxicity" or "biological relevance").
- Each property must describe a quantifiable characteristic such as molecular weight, charge, hydrophobicity, or known structural alerts.
- Only return a numbered list of five unique molecular property names.

Return the list only. Do not include explanations or SMILES fragments.

SUBSTRUCTURE MINING PROMPT:

Your task is to identify ten distinct substructures likely to contribute to clinical toxicity (CT_TOX), from {sample} that are not present in {exclusion list}, as relevant to human adverse drug reactions in the Tox21-derived ClinTox dataset.
Common toxic mechanisms include:

- Covalent binding to nucleophilic residues (e.g., Cys, Lys, DNA bases)
- Formation of reactive metabolites (e.g., quinones, radicals)
- Mitochondrial disruption or oxidative stress
- Immune system activation or hypersensitivity

Requirements:

- Return 10 strict, unique, and chemically valid SMILES substructures.
- Each must represent a plausible toxicophore or clinically relevant toxic fragment linked to CT_TOX activation.
- Format each as a valid SMILES string parsable by RDKit via `Chem.MolFromSmiles()`.
- Focus on fragments implicated in systemic toxicity, organ injury, or bioactivation in clinical pharmacology.

Return only a numbered list of 10 SMILES fragments. Do not provide explanations.

TASK FEATURE MINING OUTPUT:

1. TPSA
2. MolecularWeight
3. AromaticRings
4. LogP
5. LogS

SUBSTRUCTURE MINING OUTPUT:

1. C(=O)Cl

2. C=C(C=O)C
3. C#N
4. N(CCO)C
5. C1=CC=C(C=C1)F
6. C1=CC=C(C=C1)N
7. C(=O)N
8. C=C
9. C(Cl)(Cl)Cl
10. C(=O)OC

Complete set of Substructure and Task Feature

The Table 8 lists each different substructure obtained during Substructure Mining. The same substructure may have different representations, but the Rdkit tool will unify them into STANDARF SMARTS format. The Table 9 lists each task feature and how it is calculated by the tool during Task Feature Mining.

Table 8: Substructures and their SMILES representations.

STANDARD SMARTS	Substructure
NC=O	C(=O)N
O=CO	C(=O)O
c1ccccc1	C1=CC=CC=C1
O=S=O	[S](=O)(=O)
FC(F)F	C(F)(F)F
CBr	C(Br)
CCl	C(Cl)
C1CCNCC1	N1CCCCC1
C#N	C#N
NCCO	N(CCO)
COC=O	C(=O)OC
Clc1ccccc1	Clc1ccccc1
O=[N+][O-]	[N+](=O)[O-]
C1CCCCC1	C1CCCCC1
Fc1ccccc1	C1=CC=C(C=C1)F
CC(=O)O	OC(=O)C
CCOC=O	C(=O)OCC
O=CCl	C(=O)Cl
C1CNCCN1	N1CCNCC1
CC=O	C(=O)C
[N]	[N@]
C1CCNC1	C1CCCN1
CNC=O	C(=O)NC
NCCNC=O	C(=O)NCCN
C=C(C)C=O	C=C(C=O)C
CCC=O	C(=O)CC
C=CC=O	C(=O)C=C
CCOC(=O)OC	C(OC(=O)OCC)
CCOC(=O)CC	C(COC(=O)CC)
OCCl	C(Cl)O
C=C	C=C
Brc1ccccc1	C1=C(Br)C=CC=C1
c1ccc2ccccc2c1	C1=CC2=CC=CC=C2C=C1
Nc1ccccc1	C1=CC=CC=C1N
NC(=O)c1ccccc1	C1=C(C=CC=C1)C(=O)N
CS(=O)(=O)c1ccccc1	C1=CC=CC=C1[S](=O)(=O)C

Table 9: Task Features and Their RDKit-Based Calculation Methods

Task Feature	Calculation Method (RDKit)
MolecularWeight	`Descriptors.MolWt(mol)`
LogP	`Crippen.MolLogP(mol)`
TPSA	`rdMolDescriptors.CalcTPSA(mol)`
NumHDonors	`Lipinski.NumHDonors(mol)`
NumHAcceptors	`Lipinski.NumHAcceptors(mol)`
FormalCharge	`Chem.GetFormalCharge(mol)`
NumRotatableBonds	`Lipinski.NumRotatableBonds(mol)`
HeavyAtomCount	`mol.GetNumHeavyAtoms()`
AromaticRings	`aromatic_and_hetero_ring_counts(mol)[0]`
HeteroRings	`aromatic_and_hetero_ring_counts(mol)[1]`
NumRings	`mol.GetRingInfo().NumRings()`
C_count, N_count, O_count, ...	`Counter(atom.GetAtomicNum())`
MaxGasteigerCharge	`max(_GasteigerCharge per atom)`
MinGasteigerCharge	`min(_GasteigerCharge per atom)`
TPSA_per_MW	`TPSA / MolecularWeight`
LogP_per_MW	`LogP / MolecularWeight`
AromaticAtomRatio	`aromatic_atoms / total_atoms`
HeteroAtomRatio	`hetero_atoms / total_atoms`
BertzCT	`Descriptors.BertzCT(mol)`
FractionCSP3	`Descriptors.FractionCSP3(mol)`
Nitro_count	`mol.GetSubstructMatches('[N+](=O)[O-]')`
Aldehyde_count	`mol.GetSubstructMatches('[CX3H1](=O)[#6]')`
Michael_acceptor_count	`mol.GetSubstructMatches('O=CC=CC')`
Epoxide_count	`mol.GetSubstructMatches('C1OC1')`
LogS	`−0.8*LogP −0.01*MolWt +0.5*NumHDonors −0.01*TPSA`

G ABBREVIATIONS USED

Table 10: ABBREVIATIONS USED

Full Name	Abbreviation
NumRotatableBonds	RotB
HeavyAtoms	HeavyAtm
MolRefractivity	MolRef
NumHAcceptors	HBA
NumSecondaryAlcohols	2°OH
Hydroxylethylamine_core	HEA
Morpholine	Morph
MurckoRingAtoms	MRingA
sp3_C	sp3C
Sulfonamide	SO2NH
HeteroAtomRatio	HAR
Carboxyl	COOH
HallKierAlpha	HKAlpha
Kappa1	Kappa1
Fluoro_arene	ArF
Amide_group	CONH
NumTertiaryAmines	3°N
MolecularWeight	MolWt
FractionCSP3	FrCSP3

Table 11: Mapping of MedDRA Terms to Abbreviations

Full Name	**Abbreviation**
Hepatobiliary disorders	Hepato
Metabolism and nutrition disorders	Metabolism
Product issues	Product
Eye disorders	Eye
Investigations	Investigation
Musculoskeletal and connective tissue disorders	Muscle
Gastrointestinal disorders	Gastro
Social circumstances	Social
Immune system disorders	Immune
Reproductive system and breast disorders	Reproductive
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	Neoplasm
General disorders and administration site conditions	General
Endocrine disorders	Endocrine
Surgical and medical procedures	Surgical
Vascular disorders	Vascular
Blood and lymphatic system disorders	Blood
Skin and subcutaneous tissue disorders	Skin
Congenital, familial and genetic disorders	Congenital
Infections and infestations	Infections
Respiratory, thoracic and mediastinal disorders	Respiratory
Psychiatric disorders	Psychiatric
Renal and urinary disorders	Renal
Pregnancy, puerperium and perinatal conditions	Pregnancy
Ear and labyrinth disorders	Ear
Cardiac disorders	Cardiac
Nervous system disorders	Nervous
Injury, poisoning and procedural complications	Injury