

MCLE-Mol: Empowering LLM with Molecular Comprehension and Low-Cost Continual Evolution for Interpretable Property Prediction

Zhili Pu^{◇*}, Lantian Zhang^{◇*}, Hao Duan[◇], Zhixing Zhang[◇], Keyun Zhu[◇],
Yongqi Fan[◇], Ruihui Hou[◇], Tong Ruan^{◇†}, Yun Tang^{◇†}

[◇]East China University of Science and Technology, Shanghai, China,

Abstract

Large language models (LLMs) offer a new paradigm for molecular property prediction (MPP), yet a semantic gap between natural language and molecular representations limits LLMs’ ability to capture structure–activity relationships (SAR). Recent approaches have explored injecting structure-level information into LLMs, primarily by modeling associations based on statistical regularities. However, these methods are prone to misinterpreting coincidental associations as general principles, imposing a bottleneck on predictive performance. To tackle the challenges above, we propose MCLE-Mol, an ML–LLM–Rule collaborative framework for MPP. It bridges the semantic gap by injecting ML-derived substructure attribution values into LLMs, utilizing Context-Calibrated Substructure Attribution Rules (CCSAR) to calibrate these attributions under specific chemical contexts to mitigate such misinterpretation. In addition, MCLE-Mol introduces a low-cost continual evolution strategy that updates CCSAR with frozen model parameters to adapt to dynamic chemical spaces. Experiments on multiple benchmark datasets demonstrate that MCLE-Mol outperforms all baselines, successfully resolving the trade-off between predictive performance and interpretability. Our code is available at <https://github.com/DoorKickers/MCLE-Mol>.

1 Introduction

Molecular Property Prediction (MPP) is critical for AI-aided drug discovery (AIDD), accelerating screening and reducing experimental burdens (Chen et al., 2018; Lin et al., 2025; Vamathevan et al., 2019; Wu et al., 2018). Although Graph Neural Networks (GNNs) achieve remarkable predictive performance, their inherent black-box nature (Figure 1, top-left) lacks the interpretability

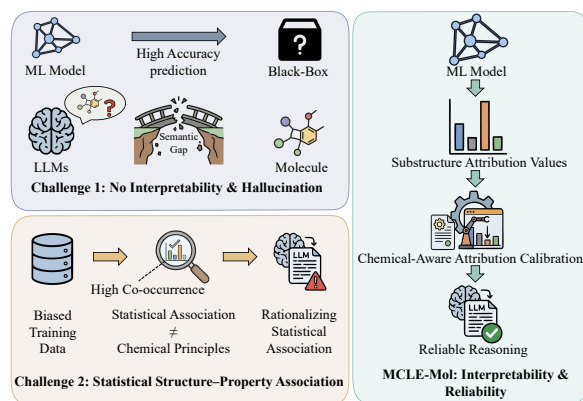


Figure 1: Motivation of MCLE-Mol. Top-Left: The performance-interpretability dilemma. Bottom-Left: Risk of misinterpreting statistical associations as chemical principles. Right: MCLE-Mol achieves reliable reasoning by injecting ML-derived attributions and applying chemical-aware calibration.

essential for rational drug design (Ying et al., 2021; Yang et al., 2019; Zhou et al., 2023b). In contrast, large language models (LLMs) offer a potential pathway toward interpretable reasoning. However, a pronounced semantic gap between natural language and molecular representations hinders LLMs from capturing structure–activity relationships (SAR) (Guo et al.; Cherkasov et al., 2014; Edwards et al., 2022), leading to hallucinated reasoning. Consequently, a dilemma persists: high-performance models lack interpretability, while LLM-based approaches—despite their potential for interpretability—typically suffer from inferior predictive performance and reliability.

Grounded in SAR principles attributing macroscopic properties to local substructures (Bon et al., 2022), recent approaches have attempted to bridge the semantic gap by injecting structural information into LLMs. These approaches can be broadly categorized into two lines of work: one directly feeds molecular structural information into LLMs to perform end-to-end property prediction, with-

* Co-first authors

† Corresponding authors.

out providing interpretability (Cao et al., 2025; Liu et al., 2023b,a); the other leverages known structure–property associations to support reasoning and provide interpretability (Xian et al., 2025). However, both paradigms typically overlook the specific chemical context of the target molecule. As shown in Figure 1 (bottom-left), relying on statistical co-occurrence patterns leads to misinterpreting coincidental associations as general chemical principles. Such statistically biased information is then rationalized by LLMs as plausible reasoning, resulting in degraded predictive performance and unreliable interpretations.

To address these issues, we propose an ML–LLM–Rule collaborative reasoning framework. As illustrated in Figure 1 (right), the framework uses substructure attribution values produced by an ML model as explicit structural cues to guide the LLM to focus on key pharmacophores. Building on this, we introduce Context-Calibrated Substructure Attribution Rules (CCSAR) to dynamically calibrate attribution values based on the specific chemical context of the target molecule. CCSAR mitigates biases caused by coincidental structure–property associations and guides the LLM to reason over more reliable structural information. Thus, our framework resolves the trade-off between predictive performance and interpretability. Moreover, since real-world chemical space evolves continuously, traditional approaches that rely on frequent retraining are computationally expensive and prone to catastrophic forgetting (Kirkpatrick et al., 2017). To overcome this bottleneck, we propose a low-cost continual evolution strategy that updates CCSAR while keeping model parameters frozen, enabling efficient adaptation to dynamic chemical spaces.

In summary, our contributions can be summarized as follows: (1) **ML–LLM–Rule Collaborative Framework**. We construct a collaborative framework that enables the LLM to reason under the specific chemical context of the target molecule, successfully resolving the trade-off between predictive performance and interpretability. (2) **Low-Cost Continual Evolution**. We propose a continual evolution strategy that updates CCSAR with frozen model parameters, enabling efficient adaptation to dynamic chemical spaces. (3) **Superior Performance and Interpretability**. Experimental results on four benchmarks (BACE, BBBP, HIV, Tox21) demonstrate that our framework achieves performance comparable to task-specific specialist

models, while maintaining inherent interpretability.

2 Related Work

Machine Learning for Molecular Property Prediction. In recent years, GNNs have become a dominant paradigm for molecular representation learning (Gilmer et al., 2017; Yang et al., 2019). By modeling molecules as topological graphs and aggregating neighborhood information of atoms and bonds through message-passing mechanisms, representative models such as DMP (Zhu et al., 2023), Graphormer-p (Ying et al., 2021), and GEM (Fang et al., 2022) have achieved competitive performance on MPP tasks. Despite their remarkable predictive performance, GNN-based models are inherently black-box in nature, which limits the transparency and interpretability of their decision-making processes (Jiménez-Luna et al., 2020). In contrast, MCLE-Mol provides interpretable predictions while maintaining performance comparable to mainstream GNN models.

LLMs for Molecular Property Prediction. In molecular property prediction tasks, recent studies have explored injecting structural information into large language models (LLMs) to enhance their comprehension of molecular structures. Mainstream approaches include introducing molecular graph features through multimodal alignment (Cao et al., 2025; Liu et al., 2023b; Lee et al., 2025), or explicitly injecting structural examples of similar molecules via retrieval-augmented generation (RAG) (Xian et al., 2025; Liu et al., 2024b). Although these methods partially capture SAR, their inference processes primarily rely on statistical regularities learned from data. This reliance leads to misinterpreting high co-occurrence patterns as general chemical principles and overlooking specific chemical contexts of the target molecule, ultimately limiting predictive capability (Stumpfe and Bajorath, 2012). A closely related line of work explores incorporating chemically grounded knowledge or rule-like guidance into molecular prediction. For example, LLM4SD (Zheng et al., 2025) leverages LLMs to synthesize literature knowledge and infer data-driven scientific knowledge, which are then transformed into interpretable feature vectors for downstream prediction with interpretable models such as random forest. M²LLM (Ju et al., 2025) incorporates molecular structure, task information, and molecular rules into a multi-view representation learning framework, where the rule view is

translated into features and fused with other views for downstream prediction. Compared with these approaches, MCLE-Mol directly integrates molecular descriptions, ML-derived substructure attributions, and retrieved chemical rules into a unified reasoning framework, while further calibrating attribution signals under the specific chemical context of the target molecule through CCSAR.

Training-Free Models. To avoid the high computational cost of continual parameter updates (Zhou et al., 2023a), training-free paradigms have emerged as a low-cost approach to improving model performance by updating external experiences instead of updating model parameters. Prior studies, such as Training-Free GRPO (Cai et al., 2025a) and FLEX (Cai et al., 2025b) demonstrate that continual evolution is possible solely through updating external experiences. Inspired by these works, MCLE-Mol dynamically maintains CCSAR to adapt to an evolving chemical space with low computational cost.

3 Methodology

As illustrated in Figure 2, we propose an interpretable framework for molecular property prediction that conducts mechanism-driven reasoning grounded in the specific chemical context of the target molecule. Furthermore, our framework enables low-cost continual evolution in dynamic chemical spaces by keeping model parameters frozen while iteratively updating the CCSAR.

3.1 Problem Definition

We formulate the task of interpretable molecular property prediction as a generation problem that integrates substructure attribution with context-aware calibration. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ denote a molecular dataset, where x represents a molecule and y is the target property label. We characterize each input molecule x as a composite feature tuple (S_x, W_x, D_x) . Here, S_x and W_x denote the set of substructures and their corresponding heuristic attribution values to the target property, respectively. D_x represents multi-view textual descriptions, specifically comprising the SMILES string, graph connectivity text, and spatial conformation text of the molecule.

We incorporate the CCSAR repository, denoted as \mathcal{R} . For a specific input x , a relevant subset of rules $\mathcal{C}_x \subset \mathcal{R}$ is retrieved to serve as context. Consequently, our objective is to learn an infer-

ence function f_θ that jointly generates the predicted value \hat{y} and the explanation text e , conditioned on the molecular features and the retrieved rules:

$$(\hat{y}, e) = f_\theta(x, S_x, W_x, D_x, \mathcal{C}_x). \quad (1)$$

3.2 Preparing Training Data

3.2.1 Computing Substructure Attributions

To explicitly reveal the impact of substructures on properties and guide the LLM’s reasoning, we introduce substructure attribution derived from an RGCN model as a heuristic signal. Specifically, we employ the Substructure-Mask Explanation (SME) method (Wu et al., 2023) to quantify the attribution of a specific substructure to the target property. SME calculates the attribution by measuring the prediction shift caused by masking the atomic features of the target substructure.

Let $f(\cdot)$ denote the RGCN prediction function, G be the molecular graph, and S be a specific substructure within G . The attribution value $Attr(S)$ is defined as:

$$Attr(S) = f(G) - f(G_{\setminus S}), \quad (2)$$

where $G_{\setminus S}$ represents the masked molecular graph with feature vectors of atoms in S set to zero.

3.2.2 Generating Molecular Descriptions

To enhance the molecular comprehension of LLMs, we develop a pipeline to generate multi-dimensional textual descriptions. Specifically, we first employ chemical tools such as RDKit (Landrums, 2006) to compute the ground-truth physicochemical properties of each molecule, including lipophilicity (LogP), topological polar surface area (TPSA), and 3D shape descriptors (e.g., NPR1 and NPR2). These attributes serve as “physical anchors” and are fed into a teacher model, such as GPT-4o (Hurst et al., 2024).

The teacher model generates textual descriptions across three dimensions: 1D SMILES sequences, 2D graph topology, and 3D spatial conformations. To ensure scientific rigor, we impose strict prompting constraints that ground the descriptions in the provided physicochemical properties. This strategy effectively mitigates hallucinations by aligning the generated text with the molecule’s physical anchors. Detailed prompt templates are provided in Appendix C.

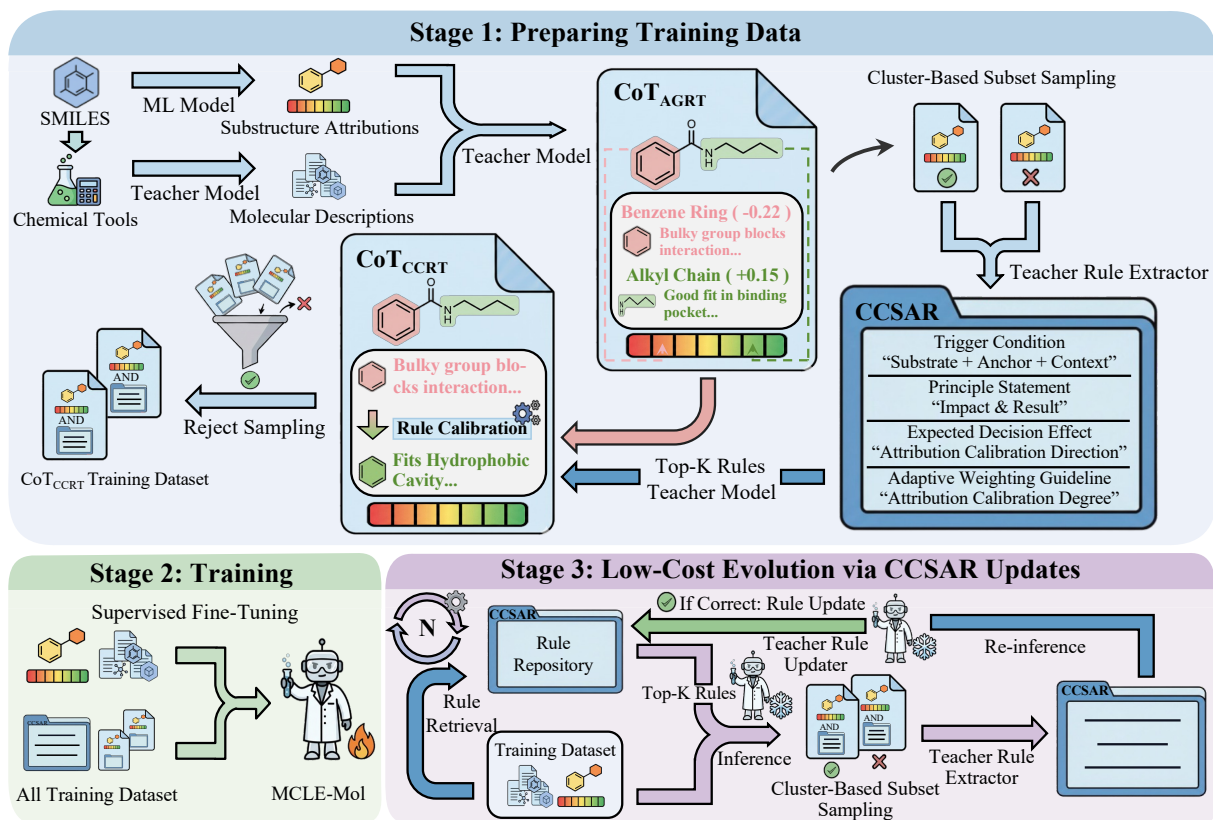


Figure 2: Overall Framework of MCLE-Mol. MCLE-Mol first executes Stage 1: Training Data Preparation to construct the CCRT dataset by incorporating ML-derived attributions and calibrating them within specific chemical contexts, proceeds to Stage 2: Model Training to optimize the LLM via supervised fine-tuning on CCRT, and achieves Stage 3: Low-Cost Evolution by solely updating the CCSAR repository.

3.2.3 Constructing AGRT

To bridge the semantic gap between natural language and molecular representations, we introduce **Attribution-Guided Reasoning Traces (AGRT)**.

Specifically, we integrate substructure attribution values into the prompt as an “explicit cue” regarding property prediction. This cue does not provide a direct answer but serves as a heuristic to stimulate the teacher model to focus on structural units with the most significant impact on the property. Guided by these cues, the model articulates the underlying chemical logic, constructing a reasoning trace that connects specific substructures to the target property. These traces are then utilized by subsequent procedures. The prompt templates used for generating AGRT are provided in Appendix D.

3.2.4 Building CCSAR

To mitigate the risk of misinterpreting coincidental structure–property associations derived from statistical regularities as general principles, we constructed an external rule repository named **Context-Calibrated Substructure Attribution Rules (CC-**

SAR). Serving as empirical heuristics, the core objective of CCSAR was to dynamically calibrate unreasonable substructure attribution values. CCSAR incorporates a “3W” context-trigger mechanism to assess the validity of attribution values within the specific chemical context. This mechanism evaluates “Who” (the substructure itself), “Where” (topological position), and “With Whom” (surrounding chemical environment) to determine if the attribution is chemically plausible.

A complete CCSAR rule contained the following key fields:

- **Trigger Condition:** Following the “3W” principle, this defined a chemical scenario of “specific substructure + specific position + specific environment.” The rule was activated only when a molecule satisfied this three-dimensional condition.
- **Principle Statement:** This explained the chemical mechanism within the given scenario.
- **Expected Decision Effect:** This guided the

direction of attribution value calibration.

- **Adaptive Calibration Guideline:** This guided the magnitude of attribution value calibration.

To extract CCSAR rules, we first divided AGRT samples into correct and incorrect reasoning samples based on predictive accuracy, and clustered both types of samples separately based on Morgan fingerprint similarity (Rogers and Hahn, 2010) into N clusters. Subsequently, we selected M representative reasoning samples from each cluster to form a sample set for the teacher model to extract CCSAR rules. Inspired by the “knowledge conflict” theory (Longpre et al., 2021; Kandpal et al., 2023), we intentionally bypassed the verification of predictive effectiveness (i.e., whether the rules can assist the model in making correct predictions). This strategy aimed to counteract the “parameter memory bias” prevalent in LLMs, where models often favor general statistical associations (e.g., nitro groups usually cause toxicity) from pre-training rather than context-specific evidence. By leveraging these rules as external constraints that effectively challenge such internal biases, we compelled the model to overcome memorized shortcuts and acquire the ability to prioritize context-based instructions for conditional reasoning. For newly extracted rules, we employed the BGE model (Xiao et al., 2024) to retrieve the Top- k semantically similar entries from the existing CCSAR repository based on the trigger conditions field, forming a reference context. The teacher model then evaluated each candidate rule against this context to determine the update strategy: candidates providing novel information were added, those capable of refining current entries were merged, and rules already semantically covered were discarded as redundant. The prompt templates for rule extraction and update decisions are provided in Appendix E.

3.2.5 Constructing CCRT

To calibrate potential biases in the attribution values of the initial AGRT, we constructed the **CCSAR-Calibrated Attribution Reasoning Traces (CCRT)**.

To accurately match the general rules in CCSAR to specific molecules in the training set, we designed a Hypothesis-Bridged retrieval mechanism. This mechanism was inspired by the Hypothetical Document Embeddings (HyDE) method (Gao et al., 2023) and aimed to bridge the semantic gap be-

tween the molecular data modality and the rule text modality. Specifically, we did not directly use raw molecular features for retrieval, but instead built a semantic “bridge” through the teacher model: First, the teacher model identifies key substructures strongly correlated with the prediction task (e.g., BBBP permeability). These substructural states are then translated into natural language descriptions following the “3W principle”. Then, the resulting description serves as the retrieval Query, while the Trigger Condition field of the rules in the CCSAR repository acts as the Key. By calculating the semantic similarity between the Query and the Key, we accurately retrieved the Top- k rules. Acting as explicit calibration directives, these rules guide the teacher model to refine the AGRT, ultimately generating the final CCRT. The corresponding prompt templates are detailed in Appendix F

3.3 Training with CCRT

We performed supervised fine-tuning (SFT) on the student model using the CCRT dataset. This training aimed to equip the model with a chemical-aware calibration capability, enabling it to dynamically calibrate substructure attributions and weigh competing factors. This training ensures that the model can adapt to an evolving chemical space solely by updating the CCSAR repository.

3.4 Low-Cost Evolution via CCSAR Updates

Real-world chemical space is dynamic. With the continual emergence of novel molecular patterns and scenarios, traditional models typically require substantial computational costs for retraining to cover the shifting data distribution. To address these challenges, we proposed a low-cost evolution mechanism based on iterative refinement of CCSAR. During this phase, the student model parameters remained frozen as a stable inference engine. We performed N rounds of inference on the training set to extract higher-quality rules and reconstruct a CCSAR capable of adapting to the shifting data distribution.

Specifically, in each iteration, the model generated reasoning traces based on the current CCSAR. Following the procedures in Section 3.2.4, we extracted candidate rules and performed incremental updates. Unlike Section 3.2.4, an additional effectiveness validation step was introduced to ensure rule quality, where each candidate rule was backfilled into the context of its source molecule for re-inference. A candidate rule proceeded to

Table 1: Main experimental results comparing MCLE-Mol with task-specific specialist models and LLM-based generalist models. The best results across all methods are highlighted in **bold**, while † denotes the best performance within the LLM-based generalist category.

Model	Method	Interpretability	BACE	BBBP	HIV	Tox21	Avg.
<i># Task-specific specialist models</i>							
GNNs	Graphormer	✗	0.7760	0.7015	0.7452	0.7589	0.7454
	Graphormer-p	✗	0.8575	0.7163	0.7788	0.7729	0.7813
	DMP(TF+GNN)	✗	0.8940	0.7780	0.8140	0.7910	0.8192
	GEM	✗	0.8560	0.7240	0.8060	0.7810	0.7917
-	Uni-Mol	✗	0.8570	0.7290	0.8080	0.7960	0.7975
-	MolXPT	✗	0.8840	0.8000	0.7810	0.7710	0.8090
-	GIT-Mol(G+S)	✗	0.8108	0.7390	-	0.7590	0.7696
-	MolCA (1D+2D)	✗	0.7980	0.7000	-	0.7720	0.7566
<i># LLM-based generalist models</i>							
-	Mol-LLM	✗	0.8080	0.8430	0.7650	-	0.8053
-	InstructMol-GS	✗	0.8210	0.7240	0.6890	-	0.7446
-	MolecularGPT(2-Shot)	✗	0.7218	0.7260	0.7204	0.6573	0.7063
-	M ² LLM	✗	0.7820	0.7550	0.7950	-	0.7773
-	MolRAG	✓	0.7225	0.5411	0.6436	0.6393	0.6366
Qwen3-8b	MCLE-Mol (Ours)	✓	0.9178	0.8581	0.7668	0.7596 [†]	0.8255

the incremental update stage only if it successfully guided the model to a correct prediction.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate MCLE-Mol on four MoleculeNet benchmarks: BACE, BBBP, HIV, and Tox21. All experiments strictly follow the official data splitting strategies recommended by MoleculeNet to construct training and test sets (Wu et al., 2018). For the larger datasets (HIV and Tox21), we use a sub-training set constructed via cluster-based sampling, with fewer than 1,000 samples selected from the original training sets.

Baselines. We categorize the compared methods into two groups. (1) Task-specific specialist models, where each model is trained and evaluated for a single molecular task, including Uni-Mol (Zhou et al., 2023b), MolXPT (Liu et al., 2023a), Graphormer (Ying et al., 2021), Graphormer-p (Ying et al., 2021), DMP (Zhu et al., 2023), GEM (Fang et al., 2022), GIT-Mol (Liu et al., 2024a), and MolCA (Liu et al., 2023b). (2) LLM-based generalist models, where LLM-based frameworks are developed for molecular prediction across multiple tasks, including Mol-LLM (Lee et al., 2025), InstructMol (Cao et al., 2025), MolecularGPT (Liu et al., 2024b), MolRAG (Xian et al., 2025), and M²LLM (Ju et al., 2025).

Metric. We use ROC-AUC (Area Under the Receiver Operating Characteristic Curve) as the evaluation metric.

Implementation details. We employed Qwen3-8B (Yang et al., 2025) as the backbone model and conducted supervised fine-tuning with LoRA (Hu et al., 2022) using the SWIFT framework (ModelScope, 2023). The training hyperparameters were configured with a learning rate of 1e-4, a warmup ratio of 0.05, a batch size of 1, and 10 training epochs. During inference, we set the temperature to 0.7 and top-p to 0.9. For each prediction, the top-k retrieved rules from CCSAR were incorporated as context, with k set to 5. For the continual evolution phase, the iterative refinement process was conducted for 3 rounds ($N = 3$). All experiments were implemented on four NVIDIA A100 (40G) GPUs. Further details are provided in Appendix A

4.2 Main Results

As shown in Table 1, we systematically compared MCLE-Mol against task-specific specialist models and LLM-based generalist models, adopting the performance at the third iteration. The results demonstrate that MCLE-Mol effectively bridges the semantic gap between natural language and molecular representations by enabling the LLM to perform reasoning anchored in the specific chemical context of the target molecule. This capability allows MCLE-Mol to capture subtle

Table 2: Ablation study evaluating the contribution of each component in MCLE-Mol. “w/o” denotes the removal of the corresponding module (ML Attribution, CCSAR, or Iterative Evolution).

Method Variant	Components			Full-Test			
	ML Attr.	CCSAR	Iter.	BACE	BBBP	HIV	Tox21
MCLE-Mol (Full)	✓	✓	✓	0.9178	0.8581	0.7668	0.7596
w/o Iterative Evolution	✓	✓	✗	0.8816	0.8081	0.7014	0.7506
w/o CCSAR	✓	✗	✗	0.7074	0.7729	0.6217	0.7171
w/o ML Attribution	✗	✗	✗	0.5134	0.5827	0.5261	0.5968

structure-property relationships, thereby significantly outperforming existing generalist models across all datasets and surpassing optimized specialist models on BACE and BBBP. Ultimately, MCLE-Mol achieves the highest average ROC-AUC of 82.55%. Notably, even with fewer than 1,000 samples for fine-tuning on the large-scale HIV and Tox21 datasets, MCLE-Mol still achieved competitive performance on these two benchmarks compared to baselines trained on full data, confirming that MCLE-Mol significantly reduces the reliance on large-scale annotated data. Furthermore, unlike most black-box models that only output binary predictions, MCLE-Mol generates both high-performance predictions and natural language interpretations based on substructure attribution, successfully resolving the trade-off between predictive performance and interpretability to provide trustworthy decision support for drug discovery.

4.3 Ablation Study

To evaluate the contribution of each key component in MCLE-Mol, we perform an ablation study on the official full test sets. As shown in Table 2, we construct three variants by progressively removing components and compare them with the full model (MCLE-Mol Full). The full model achieves the best performance across all datasets, indicating that all components are necessary and complementary. The detailed analysis is as follows:

- **Effect of ML Attribution:** Compared with the variant that removes all components, introducing substructure attribution significantly improves performance. This indicates that using attribution to guide the model to focus on key molecular substructures is necessary.
- **Effect of the CCSAR:** Adding the CCSAR module on top of ML Attribution further improves performance. This suggests that ML attribution values contain statistical bias, and CCSAR can correct unreasonable attributions based on the specific chemical context of the

target molecule, making the reasoning process more consistent with chemical principles.

- **Effect of iterative evolution:** Adding an iterative mechanism yields the best performance, demonstrating that this mechanism enables the model to evolve continually within a dynamic chemical space.

Overall, all components are complementary and jointly contribute to MCLE-Mol’s performance.

4.4 CCSAR Generalization

To evaluate the generalizability of CCSAR to out-of-distribution molecules, we measured the rule trigger rate on the test set. A trigger event was recorded when a retrieved rule was utilized during inference. Quantitatively, the trigger rates exceeded 90% on most datasets (specifically, 98.7% for BACE, 91.9% for BBBP, and 90.4% for Tox21, with HIV at 88.5%), indicating that CCSAR generalizes well and effectively covers out-of-distribution samples. Furthermore, to assess the transferability of CCSAR as an independent module, we applied it to distinct models with different parameter scales. Since CCSAR is designed to calibrate ML attribution values, we conducted controlled ablation experiments to disentangle the contributions of ML attributions and CCSAR. As shown in Figure 3, introducing ML attributions alone (middle bar) already improved performance compared to the original inference (left bar), confirming the effectiveness of substructural attribution information. Building on this, adding CCSAR (right bar) further improved performance across all models, indicating that CCSAR acts as an effective, model-agnostic plugin that yields improvements regardless of model architecture or parameter scale.

4.5 Evolution Analysis

As illustrated in Figure 4, we analyzed the evolution process by monitoring the performance changes and the number of newly extracted CCSAR rules across three iterations. During the initial

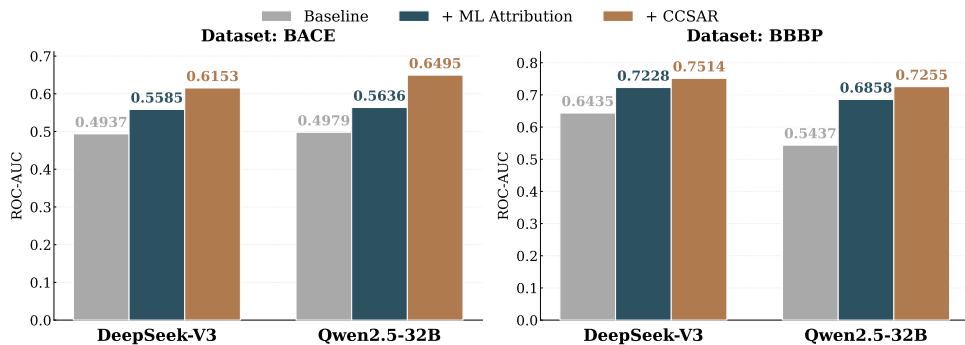


Figure 3: ROC-AUC comparison on BACE and BBBP datasets using DeepSeek-V3 and Qwen2.5-32B across three settings: Baseline, +ML Attribution, and +CCSAR.

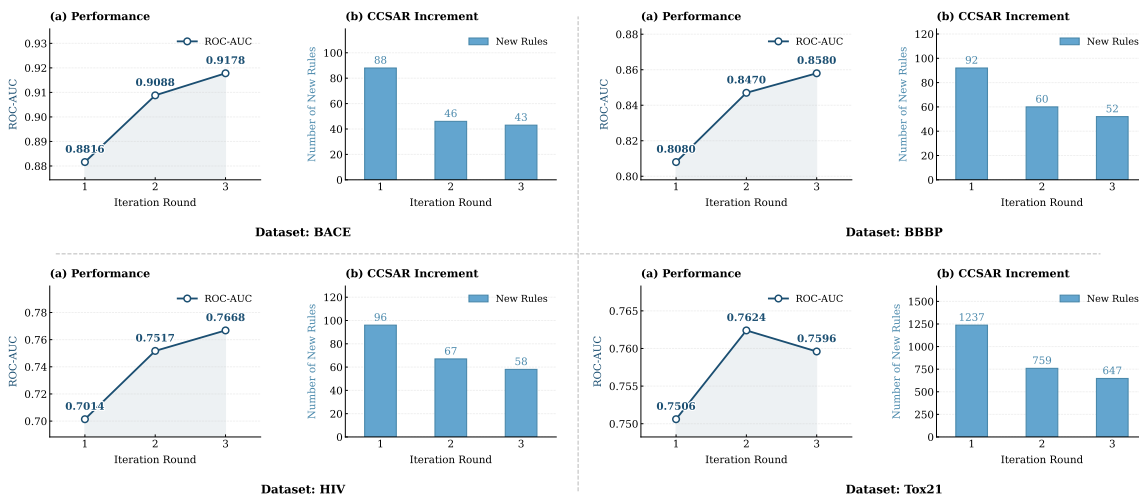


Figure 4: Evolution analysis across four datasets. (a) The trajectory of predictive performance (ROC-AUC) over three iterations. (b) The number of newly extracted CCSAR rules in each corresponding iteration.

phase, the CCSAR repository expanded rapidly; however, as iterations progressed, the number of new rules diminished, while the model performance approached a plateau after an initial rapid rise. Notably, the Tox21 dataset exhibited a distinct early convergence pattern, implying that the model efficiently acquired the majority of key knowledge within the first iteration and focused on solving long-tail samples in subsequent rounds. This trend indicates that MCLE-Mol efficiently captured key SAR patterns in the early stages, and subsequently minimized rule redundancy while progressively addressing long-tail samples.

4.6 Impact of the Number of Retrieved Rules

To determine the optimal number of retrieved rules, we conducted a sensitivity analysis on the parameter $k \in \{3, 5, 7, 9\}$ using the BACE and BBBP datasets. As shown in Table 3, model performance improved significantly as k increased from 3 to 5, which indicates that providing a sufficient amount of rule context can effectively support the model’s

reasoning process. However, a slight performance decline was observed when k was further increased to 7 or 9. This suggests that introducing too many retrieved rules may bring in irrelevant or redundant information, which can distract the model and interfere with its attention mechanism. In addition, larger k values substantially increase the input length, resulting in higher inference costs and reduced computational efficiency. Balancing predictive performance and computational efficiency, we selected $k = 5$ as the default setting.

Table 3: Sensitivity analysis of the retrieval parameter k . We investigate the impact of utilizing different numbers of top- k rules retrieved from the CCSAR repository as contextual input.

Top- k	BACE	BBBP	Avg.
3	0.9067	0.8238	0.8653
5 (Default)	0.9178	0.8581	0.8880
7	0.9103	0.8502	0.8803
9	0.9113	0.8382	0.8748

5 Conclusion

In this paper, we present MCLE-Mol, an interpretable framework that balances predictive performance and transparency in molecular property prediction. By bridging the semantic gap with ML attributions and mitigating statistical biases via the CCSAR repository, MCLE-Mol enables LLMs to reason within specific chemical contexts, ensuring reliable interpretations. Our iterative evolution strategy allows for low-cost adaptation to dynamic chemical spaces by updating the repository without retraining. Extensive experiments validate that MCLE-Mol achieves superior performance and high rule trigger rates, effectively capturing universal SARs and addressing long-tail samples.

Limitations

Despite the promising results, MCLE-Mol presents two primary limitations. First, the framework relies on the quality of the teacher model and external chemical tools; consequently, potential hallucinations or extraction errors may propagate to the student model. Second, although MCLE-Mol incorporates 3D descriptors, it does not yet explicitly model the complete 3D geometric graph structure, which constrains its capability in distinguishing complex stereoisomers.

Acknowledgments

We thank the anonymous reviewers and the area chair for their valuable feedback and constructive suggestions, which helped improve the clarity and overall presentation of this paper. This work was supported by the National Key R&D Program of China under the Key Special Project "Bio-Information Integration (BT and IT Integration)" (Project No. 2023YFF1204904).

Ethics Consideration

This study exclusively utilized publicly accessible datasets for all experiments, thereby precluding any issues regarding data privacy or confidentiality. We have also rigorously assessed the broader implications and potential deployment scenarios to safeguard against misuse or unintended harm. Consequently, we affirm that this work strictly adheres to established ethical guidelines.

References

- Marta Bon, Alan Bilsland, Justin Bower, and Kirsten McAulay. 2022. Fragment-based drug discovery—the importance of high-quality molecule libraries. *Molecular Oncology*, 16(21):3761–3777.
- Yuzheng Cai, Siqi Cai, Yuchen Shi, Zihan Xu, Lichao Chen, Yulei Qin, Xiaoyu Tan, Gang Li, Zongyi Li, Haojia Lin, and 1 others. 2025a. Training-free group relative policy optimization. *arXiv preprint arXiv:2510.08191*.
- Zhicheng Cai, Xinyuan Guo, Yu Pei, Jiangtao Feng, Jinsong Su, Jiangjie Chen, Ya-Qin Zhang, Wei-Ying Ma, Mingxuan Wang, and Hao Zhou. 2025b. Flex: Continuous agent evolution via forward learning from experience. *arXiv preprint arXiv:2511.06449*.
- 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 *Proceedings of the 31st International Conference on Computational Linguistics*, pages 354–379.
- Hongming Chen, Ola Engkvist, Yin Hai Wang, Marcus Olivecrona, and Thomas Blaschke. 2018. The rise of deep learning in drug discovery. *Drug discovery today*, 23(6):1241–1250.
- Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, and 1 others. 2014. Qsar modeling: where have you been? where are you going to? *Journal of medicinal chemistry*, 57(12):4977–5010.
- Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*.
- Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777.
- 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.
- Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks, 2023. URL <https://arxiv.org/abs/2305.18365>.

- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. 2020. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584.
- Jiaxin Ju, Yizhen Zheng, Huan Yee Koh, Can Wang, and Shirui Pan. 2025. **M2llm: Multi-view molecular representation learning with large language models**. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 7437–7445.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pages 15696–15707. PMLR.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.
- Greg Landrum. 2006. **Rdkit: Open-source cheminformatics**. Accessed: 2026-01-04.
- 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*.
- Bangjiang Lin, Shujie Yan, and Bowen Zhen. 2025. A machine learning method for predicting molecular antimicrobial activity. *Scientific Reports*, 15(1):6559.
- Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024a. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073.
- Yuyan Liu, Sirui Ding, Sheng Zhou, Wenqi Fan, and Qiaoyu Tan. 2024b. Molcargpt: 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. *arXiv 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.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. *arXiv preprint arXiv:2109.05052*.
- ModelScope. 2023. Swift: Scalable lightweight infrastructure for fine-tuning. <https://github.com/modelscope/swift>.
- David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.
- Dagmar Stumpfe and Jurgen Bajorath. 2012. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 55(7):2932–2942.
- 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.
- 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.
- Zhenxing Wu, Jike Wang, Hongyan Du, Dejun Jiang, Yu Kang, Dan Li, Peichen Pan, Yafeng Deng, Dongsheng Cao, Chang-Yu Hsieh, and 1 others. 2023. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585.
- Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025. Molrag: unlocking the power of large language models for molecular property prediction. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15513–15531.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.
- 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*.

- Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, and 1 others. 2019. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388.
- 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? *Advances in neural information processing systems*, 34:28877–28888.
- Yizhen Zheng, Huan Yee Koh, Jiabin Ju, Can Wang, and Shirui Pan. 2025. [Large language models for scientific discovery in molecular property prediction](#). *Nature Machine Intelligence*.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.
- Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023b. Uni-mol: A universal 3d molecular representation learning framework.
- Jinhua Zhu, Yingce Xia, Lijun Wu, Shufang Xie, Wengang Zhou, Tao Qin, Houqiang Li, and Tie-Yan Liu. 2023. Dual-view molecular pre-training. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3615–3627.

A Implementation Details

In our implementation, we employ distinct teacher models to optimize performance across different stages: GPT-5-mini is utilized for constructing the AGRT and CCRT datasets, while GPT-4.1-mini is applied for rule extraction and dynamic CCSAR repository updates. Regarding the rule extraction strategy for the BBBP, BACE, and HIV datasets, we adopt a clustering-based sampling approach where samples are categorized by reasoning correctness and independently partitioned into 100 distinct clusters using Morgan fingerprint-based clustering, with a single representative selected from each. For the Tox21 dataset, given its larger scale, we construct an expanded sampling pool from the training set containing 500 samples per endpoint, while maintaining the same clustering and selection protocols as the other datasets.

B Case Study

To provide an intuitive understanding of how MCLE-Mol resolves core challenges in molecular reasoning, we present two representative case studies in Tables F1 and F2. Addressing the semantic gap (Table F1), the baseline model (GPT-4o-mini) relies on superficial structural features, hallucinating that the molecule's hydrogen bonding capacity supports favorable interactions, resulting in a False Positive. In contrast, MCLE-Mol accurately identifies that the elevated Topological Polar Surface Area (TPSA) induces high desolvation penalties and entropic costs, correctly classifying the molecule as inactive based on physical reality. Addressing context-aware reasoning (Table F2), the baseline incorrectly identifies a protonated ammonium group as a valid ionic anchor based on general statistical associations. However, MCLE-Mol leverages the CCSAR repository to recognize that within this specific chemical context, the adjacent electron-withdrawing groups and steric shielding nullify the ammonium's binding potential. This correctly reverses the prediction to negative, confirming the framework's ability to capture complex, non-additive SAR patterns that statistical baselines miss.

Table F1: Case Study on bridging the semantic gap between natural language and molecular representations.

Molecular Information
SMILES: <chem>S1(=O)(=O)CC(Cc2cc(O)c(N)c(F)c2)C(O)C([NH2+])Cc2cc(ccc2)C(C)</chem>
GPT-4.1-mini Output
<p><think></p> <p>The molecule under analysis has several features that align well with known BACE-1 inhibitors, which commonly exhibit balanced polarity, moderate molecular weight, and the capacity to form specific interactions within the enzyme's active site ... Polarity and H-bonding: The molecule has 4 hydrogen bond donors and 5 acceptors, along with a TPSA of 117.23 Å². BACE-1 inhibitors typically require a moderate polar surface area to engage in key hydrogen bonds within the catalytic aspartyl dyad and adjacent binding pockets. This molecule's hydrogen bonding capacity and polar surface area support the formation of such interactions ... The presence of two aromatic rings and a sulfur-containing aliphatic heterocycle provides a scaffold capable of engaging in π-stacking and hydrophobic interactions within the enzyme's S1 and S3 pockets ... Therefore, the molecule likely acts as a BACE-1 inhibitor.</p> <p></think></p> <p><result> Positive (Incorrect) </result></p>
MCLE-Mol Output (Ours)
<p><think></p> <p>**Knowledge Evaluation**: Triggered rules: - Knowledge 1 (Electron-withdrawing bulky halogenated aromatic substituent adjacent to central polyfunctional heterocyclic scaffold + nearby polar substituents): triggered by the fluorophenyl ring (c1cc(O)c(N)c(F)c1) being directly attached to the central sulfone-containing aliphatic heterocycle and the presence of nearby polar groups (phenolic OH, aniline NH, and the charged benzylammonium). Implication: steric/electronic interference with warhead orientation and reduced effective basicity of nearby nitrogens, increasing desolvation and entropic penalties; partial compensation possible from scaffold preorganization and hydrophobic contacts ... Fragment *C1CS(=O)(=O)CC(*)C1O*: -0.16 → central sulfone-containing aliphatic heterocycle with a ring-bound hydroxyl increases local polarity and steric bulk, which raises desolvation cost and can misalign the warhead relative to the Asp dyad, producing a moderate negative contribution. Mechanistic Refinement: Applying Knowledge 1 and 7, the central sulfone ring is adjacent to an electron-withdrawing, halogenated aromatic substituent and nearby polar groups; this arrangement is predicted to reduce the effective basicity/protonation propensity of proximal basic centers and to sterically/electronically misalign the warhead geometry with Asp32/Asp228, increasing desolvation and entropic penalties. The sulfone's rigidity can partially preorganize the scaffold, but it also enforces a polar surface that is costly to desolvate in the active site ... Property TPSA (117.23): elevated polar surface area increases desolvation cost and reduces passive CNS penetration, imposing a penalty on effective brain exposure and on net binding affinity in a CNS target context...</p> <p></think></p> <p><answer> Negative (Correct) </answer></p>

Table F2: Case Study on Effectiveness of the CCSAR module.

Molecular Information
SMILES: <chem>Fc1cc(cc(F)c1)CC(NC(=O)C)C(O)C[NH2+]C1(CCCCC1)c1nnn(c1)C12CC3CC(C1)CC(C2)C3</chem>
Model w/o CCSAR
<p><think> Fragment *c1cc(F)cc(F)c1: 0.019 → difluorophenyl gives hydrophobic π-contact with S-pockets; small positive because peripheral ... Fragment *[NH2+]*: 0.009 → protonated ammonium enables electrostatic attraction to Asp but is sterically shielded; very small positive. ... Synthesis: Polar motifs (protonated amine and secondary alcohol) form the catalytic salt-bridge/H-bond network while bridged polycycle and cyclohexyl supply hydrophobic pocket filling. The acetamide attenuates a key basic site and reduces that benefit.</p> <p></think></p> <p><result> Positive (Incorrect)</p> <p></result></p>
Model w/ CCSAR
<p><think> **Knowledge Evaluation**: Triggered rules: - Knowledge 6 (Protonated aliphatic nitrogen + central scaffold + adjacent electron-withdrawing aromatic substituent): triggered by the protonated aliphatic ammonium ([NH2+]) located near the difluorophenyl and bulky aliphatic substituents; implies steric shielding and reduced accessibility/basicity of the aliphatic warhead, diminishing optimal ionic interaction with the Asp dyad ... Fragment *[NH2+]*: 0.009 → very small positive; the protonated aliphatic ammonium is a potential ionic anchor but is sterically shielded by nearby quaternary centers and bulky aliphatic rings. Mechanistic Refinement: Applying Knowledge 6,7,10 — the protonated aliphatic nitrogen is proximal to the difluorophenyl and bulky aliphatic substituents, which sterically shield and electronically reduce its accessibility and effective basicity; this diminishes its ability to form an optimal salt-bridge with Asp32/Asp228 despite its formal positive charge. Final Rationale: The aliphatic ammonium is a weak, partially shielded ionic anchor; it contributes little to reliable catalytic dyad engagement ... **Final Verdict Justification**: Despite favorable hydrophobic and polar contacts, mechanistic rules indicate the central heterocyclic and aliphatic nitrogens are electronically and sterically compromised, reducing reliable ionic anchoring to Asp32/Asp228; this subversive penalty outweighs compensatory hydrophobic fit, so the molecule is unlikely to be an active BACE1 inhibitor.</p> <p></think></p> <p><answer> Negative (Correct)</p> <p></answer></p>

C Prompt Templates for Molecular Descriptions

This section presents the prompt templates used to generate multi-dimensional textual descriptions (1D, 2D, and 3D) from molecular structures. These prompts ensure that the generated text is strictly grounded in calculated physicochemical properties to serve as reliable physical anchors.

Get Molecular Descriptions Prompt

Role: You are an expert Computational Chemist and Medicinal Chemist.

Task: Analyze the provided [Molecular Fact Sheet] to generate a description focused on **Geometric Topology**, **Electronic Properties**, and **Structural Liabilities**.

OBJECTIVE: The output will be used as a structured text feature for AI models predicting:

- Bioactivity** (Target binding, Enzyme inhibition).
- ADMET** (Permeability, Solubility, Metabolism).
- Tox/Safety** (Reactivity, Off-target effects).

UNIVERSAL ANTI-HALLUCINATION RULES:

- NO SMILES RECONSTRUCTION**: NEVER attempt to generate or "guess" SMILES strings.
- Fact Adherence**: The section **[UNIVERSAL STRUCTURAL FACTS]** is the absolute ground truth. Align your description strictly with the listed Ring Counts, Sizes, and Complexity.
- Conceptual & Global Description**:
 - Since specific substituent positions are not listed in the facts, describe functional groups (e.g., Fluorine, Amines) **collectively** or **globally**.
 - Do NOT guess which specific ring bears which substituent (e.g., avoid saying "Ring A has a Fluorine" unless you are certain; instead say "The scaffold is substituted with Fluorine").

Instructions:

Output ONLY the three XML sections.

Content Requirements:

1. <SMILES>

Goal: Electronic Nature & Physicochemical Drivers

- Identity**: Classify the molecule based on [Core Scaffold Analysis] (e.g., "A molecule featuring a [Insert Core Desc] core"). Mention Formula and Charge.
- Electronic Profile**: Analyze the **collective** impact of functional groups from [Heteroatom Composition].
 - Reasoning**: Explain how the presence of electronegative atoms (F, O, N, Cl) **globally** modulates the electronic surface (polarity) or creates **reactive centers** (relevant for Tox), without assigning them to specific rings.
- ADMET**: Relate LogP and TPSA to **membrane permeability** and **solubility**.

2. <GRAPH>

Goal: Geometric Skeleton & Pharmacophore

- Ring Census**: State Ring Count and Types from facts.
- Architecture**: Describe connectivity using [Connectivity Architecture] (Isolated vs. Fused/Mixed).
- Complexity**: If [Structural Complexity] lists features like "Quaternary Carbons" or "Spiro Atoms", highlight them.
 - Reasoning**: Explain how these features introduce **rigidity** (good for binding specificity) or **metabolic stability** (blocking metabolic sites).
- Scaffold**: Describe the central framework using [Core Scaffold Analysis]. Mention that side chains extend from this core.

3. <SPACE>

Goal: 3D Shape, Selectivity & Entropy

- Shape**: Use 'Calculated Shape Class' to describe the molecular envelope.
- Complexity & Selectivity**: Interpret 'Fraction CSP3'.
 - >0.4: "Escapes flatland" -> Higher geometric complexity often improves **target selectivity** (SIDER).
 - <0.3: Planar/Aromatic dominated -> Potential for **promiscuous binding** (stacking).
- Dynamics**: Use 'Rotatable Bond Count' to discuss the **entropic cost** of binding vs. induced fit capability.

Input Data:

{molecular_context}

D Prompt Templates for Constructing AGRT

The templates for the Attribute-Guided Reasoning Tuning (AGRT) dataset guide the model to produce step-by-step reasoning traces derived from ground-truth attributes. This allows the student model to learn the fundamental mapping between structural features and properties.

Generate AGRT prompt

You are a medicinal chemistry expert analyzing whether a given molecule acts as a compound with **{...Dataset-Specific Information}**.

Your goal is to generate objective, mechanism-based reasoning that integrates:

- Global molecular physicochemical properties,
- Substructure composition and their contextual interplay,
- Quantitative contribution values derived from masking-based ML computation.

=====

BACKGROUND (contribution definition)

=====

Each contribution value is computed by masking the substructure, re-predicting the property probability with an ML model, and processing the probability change into a numeric score.

- Positive → the substructure, in this molecule's context, increases predicted **{...Dataset-Specific Information}** probability.
- Negative → the substructure, in this molecule's context, decreases predicted **{...Dataset-Specific Information}** probability.
- Magnitude → the absolute value reflects effect strength; values near zero indicate weak effect.
- These contributions are **context-dependent** (they are computed within the full molecular environment and implicitly reflect fragment interactions).
- The provided list may not cover all fragments in the molecule; treat it as a focused, partial attribution map.

IMPORTANT: **Do not** decide activity by naïvely summing these numbers. The purpose of these values is to **point** the model to which fragments to examine and to suggest relative influence; you must explain **why** each fragment's value is positive/negative and why its magnitude is as reported, in the context of the whole molecule.

=====

TASK INSTRUCTIONS

=====

You must perform structured molecular reasoning before answering.

Note: A Reference_Label is provided for orientation only to help the generator learn correct-style outputs. **Do not** mention, quote, or reveal the Reference_Label anywhere in your reasoning. The model's final decision must be justified solely by its integrated reasoning and may differ from Reference_Label.

For each substructure in VALID_SUB_SMILES (briefly, 1–2 short clauses each):

- State the reported contribution value and its sign.
- Explain, using evidence from SMILES, GRAPH_DESCRIPTION, SPACE_INFORMATION, and PROPERTY_DESCRIPTION, **why** this fragment shows that sign in this molecular context (e.g., hydrophobic aromatic systems promoting binding into relevant hydrophobic pockets of the biological target, basic amines forming salt-bridges with target active sites, polar/ionized groups reducing cell uptake, metal-chelating moieties interfering with the target mechanism, bulky polar substituents hindering target binding or cell permeation).
- Explain **why** the reported magnitude is reasonable (e.g., a large hydrophobic scaffold that complements a target pocket → large positive; a strongly ionized polar headgroup that prevents cell entry → strong negative).

Then:

- Provide a focused analysis of the **global molecular properties** in PROPERTY_DESCRIPTION, **but only comment on those properties that have a materially important effect** on potency and cell-based activity for this molecule (e.g., LogP if it affects cellular membrane permeability and intracellular exposure, TPSA if it limits cell uptake, MW if it restricts diffusion or bioavailability, pKa if it alters ionization and target interaction). For each such selected property, state how that specific value mechanistically supports or opposes {...Dataset-Specific Information}. Do not list or comment on properties that are irrelevant or have negligible impact.
- Describe how these fragment-level effects **interact** (synergy, cancellation, internal H-bonding masking polarity, conformational effects enabling target fit) and how global properties (LogP, TPSA, MW, HBD/HBA, rotatable bonds, pKa) modify or validate those fragment-level interpretations.
- Do NOT invent arbitrary numeric thresholds or external data.

Reasoning constraints:

- Inside `` be concise but explicit: each substructure explanation should be one short clause; the interaction/global synthesis may be 2-4 short sentences. Total `` length: 120-220 words.
- Do NOT use hedging/speculative words ("maybe", "could", "might", "seems", "likely", "possibly", "perhaps", "uncertain").
- Do NOT perform a raw arithmetic sum of contribution values to decide activity.

NEW OUTPUT RULE (required):

- The `` section **must end** with a single-line explicit decision in the format: `Decision: yes` or `Decision: no`
This decision line counts as part of `` and must reflect the integrated reasoning.

=====
 ### OUTPUT FORMAT (strict)
 =====

```
<think>
Per-substructure lines: "Fragment X (SMILES_or_label): contribution_value → brief explanation
of sign and magnitude."
Selective property analysis lines: "Property Y (value): brief mechanistic interpretation
supporting or opposing {...Dataset-Specific Information} (only include properties with
material impact)."
Synthesis: 2-4 short sentences integrating interactions and global properties.
Decision line: "Decision: yes" or "Decision: no"
</think>
```

```
<answer>yes</answer> or <answer>no</answer>
```

Critical rules for ``:

- `` must be exactly one token: `yes` or `no` (lowercase), matching the `Decision:` line in `` exactly.
- No additional text, punctuation, or commentary in ``.

FORBIDDEN:

- No text outside the two tags.
- No markdown, headings, notes, or external assumptions.

=====
 ### INPUT FIELDS
 =====

```
SMILES: {smiles}
SMILES_DESCRIPTION:
{smiles_description}
GRAPH_DESCRIPTION:
{graph_description}
SPACE_INFORMATION:
{space_description}
PROPERTY_DESCRIPTION:
{property}
VALID_SUB_SMILES: {valid_sub_smiles}           # A list of substructures related to
 {...Dataset-Specific Information}.
VALID_CONTRIBUTIONS: {valid_contributions}     # Numerical contribution values aligned with
 VALID_SUB_SMILES.
```

```
Reference_Label: {label} # Provided for orientation only – DO NOT mention
or quote this label in the reasoning. The final answer must be based on the model's own
integrated reasoning (it may agree or disagree with Reference_Label).
```

E Prompt Templates for Building CCSAR

E.1 Extracting CCSAR Rule

This template is designed to distill explicit chemical knowledge from representative samples into interpretable "Condition → Implication" rules, ensuring their generalizability across different chemical contexts.

Extract Rule For Correct CoT Prompt

```
## ROLE: Critical Pharmacology Scientist & Knowledge Auditor

## YOUR TASK
You are a senior medicinal chemist and pharmacology/DMPK expert specializing in
{...Dataset-Specific Information} drug discovery.
Your input is a Correctly Predicted Molecule from the {...Dataset-Specific Information}
activity task.
Your mission is to analyze its `CoT_Analysis` (reasoning process) and extract one single
key mechanism-based knowledge rule.

Important: You must focus on extracting exactly ONE key knowledge rule.

Your task is a two-stage process:

1. CRITICAL AUDIT:
Critically audit the provided `CoT_Analysis`.
Core Question: Does the model "genuinely understand" why this molecule shows (or fails to
show) {...Dataset-Specific Information}, or was it just a "Lucky Guess"?
Use your expertise in physicochemical properties, cell permeability and uptake,
target-binding, electronic distribution, and 3D conformational analysis to validate the
mechanistic soundness.

2. BRANCHED EXTRACTION:
Based on your audit, choose one of the two paths:
- [CASE A] REINFORCING Path (Sound Reasoning): The CoT correctly identified the key
mechanism driving the result. Your goal is to distill this valid logic.
- [CASE B] CORRECTIVE Path (Lucky Guess / Flawed Logic): The CoT provided shallow or
incorrect reasons (e.g., just LogP). Your goal is to uncover the real deep
mechanism and formulate a rule to fix the attribution logic.

---

## REQUIRED ANALYSIS PROCESS

### Step 1. Identify Structural & Functional Context
Determine the key substructures or physicochemical motifs.
Abstract these into functional roles (e.g., hydrophobic scaffold engaging a target
hydrophobic pocket, basic amine forming a salt bridge with an active-site residue,
metal-chelating moiety interacting with catalytic metal ions, polar handle reducing cell
uptake).

---

### Step 2. Evaluate CoT-Truth Alignment
Compare the model's successful_cot_analysis with the true label and the provided
valid_contributions:
- Analyze how the CoT correctly identified the key features, interactions, or properties (e.g.,
lipophilicity, size, efflux liability) that led to the correct label.
- Identify the most critical statement or insight in the CoT that explained the mechanism
behind these contributions (e.g., why a specific motif enhanced the desired activity).
- Summarize why the model's reasoning on these specific chemical features led to the correct
prediction.
```

Step 3. Analyze Key Success Factor (Core Diagnosis)

This is the most important step. Based on the molecule and the analysis in Step 2, diagnose the root cause of the **success** in predicting {...Dataset-Specific Information}.

You should consider the following critical factors as **guiding hints** (including but not limited to):

- **Target Recognition & Active Site Compatibility:**
(e.g., Did the model correctly identify moieties that complement the relevant biological target binding pockets? Was the hydrophobic/polar vector arrangement correctly recognized as facilitating productive engagement?)
- **Conformational Adaptability & Pharmacophore Alignment:**
(e.g., Did the molecule's flexibility or semi-rigid scaffold allow induced-fit binding? Did the model correctly factor in pharmacophore alignment for productive interactions?)

Your diagnosis should then evaluate the success against these points:

1. **Mechanistic Elucidation (Deep Reasoning):**
Did the CoT go beyond merely citing the substructure? Did it provide a specific **physicochemical or structural rationale**.
2. **Contribution Alignment (Surface Reasoning):**
Did the CoT correctly identify the dominant substructure but only provide generic justifications (e.g., "this group is lipophilic" without context)?
3. **Trivial / Non-Generalizable:**
Was the success trivial (e.g., a known scaffold) or based on a highly specific feature that cannot be generalized?

Step 4. Formulate the Decision-Triggering Knowledge Rule

Based on your diagnosis in Step 3, formulate a precise rule that captures this success using the following universal structure:

- ``principle_id``: (CRITICAL NAMING CONVENTION: Use **'DT_XX_NEW'** for a new rule).
- ``trigger_condition``:
Construct a strictly structured condition string using the exact format: **"[WHO] + [WHERE] + [WITH WHOM]"**.
- CRITICAL INDEXING CONSTRAINT (Structural Objectivity):** This field acts as a search index. It must describe **"What exists"** and **"Where it is"**, but NEVER **"How it behaves"** (State/Outcome).
- GENERALIZATION REQUIREMENT:** You must abstract the specific atoms into **representative chemical classes** to ensure the rule applies to future analogs.
 1. **[WHO] Identity:** The functional group class or core motif.
 2. **[WHERE] Topology:** The structural anchor.
 3. **[WITH WHOM] Environment:** The neighboring structural context.
- > **STATE-DESCRIPTION BAN:** You are **STRICTLY PROHIBITED** from using subjective state adjectives or outcome-based verbs in this field.
- > **NO NARROW EXAMPLES:** Do **NOT** include parenthetical examples in the trigger string.
- > **Constraint:**
 - > 1. NO NUMERICAL THRESHOLDS.
 - > 2. QUALITATIVE CONVERSION only.
 - > 3. The final output string MUST use a **plus sign ('+')** with single spaces.
 - > 4. Only include chemically necessary factors.
- ``principle_statement``:

```

A concise mechanistic rationale describing why the structural combination results in a
specific state or effectiveness.

- `expected_decision_effect`:
  Explicitly state the Direction of Adjustment for the contribution weighting and the
  Target of that adjustment.

- `adaptive_weighting_guideline`:
  Define the Magnitude of Adjustment required (minor attenuation, moderate adjustment,
  major reweighting).

---

FINAL OUTPUT FORMAT (JSON)
Output only the JSON object.
{
  "justification_of_action": "A brief, 1-2 sentence explanation of your action.",
  "extracted_knowledge": [
    {
      "principle_id": "...",
      "trigger_condition": "[WHO] + [WHERE] + [WITH WHOM]",
      "principle_statement": "...",
      "expected_decision_effect": "...",
      "adaptive_weighting_guideline": "..."
    }
  ]
}

---

INPUT DATA

SUCCESSFUL ATTEMPT DATA

* Successful CoT Analysis:
{cot_analysis}

* Correctly Predicted Molecule
* SMILES: {smiles}
* Ground Truth Label: {ground_truth_label}
* Model Prediction: {predicted_label}
* SMILES_description: {SMILES_description}
* SMILES_graph_description: {SMILES_graph_description}
* SMILES_space_description: {SMILES_space_description}
* Molecular_property: {SMILES_property}
* valid_sub_smiles: {valid_sub_smiles}
* valid_contributions: {valid_contributions}

```

Extract Rule For Incorrect CoT Prompt

```

ROLE: Decision-Triggering SAR Knowledge Extractor (for Misprediction Root Cause Analysis)

YOUR TASK
You are a senior medicinal chemist and DMPK expert performing a root-cause analysis on a
mispredicted molecule for the {...Dataset-Specific Information} activity prediction
task.

Your mission is to extract mechanism-based, decision-triggering knowledge that captures the
corrective reasoning required to fix the error.
This knowledge must explain why the model's CoT reasoning led to an incorrect decision* and
how contribution weighting should be adjusted*.

Important: You must focus on extracting exactly ONE key knowledge rule.

You will be given:
- A single molecule sample that was predicted incorrectly by the model (its CoT
  reasoning and the true ground truth label are both provided).

```

- The goal is to identify which **structural, physicochemical, or spatial reasoning factors** caused the model to misjudge the {...Dataset-Specific Information}, and to formulate **one key, generalizable decision-triggering rule** that can correct such misweighting behavior.

REQUIRED ANALYSIS PROCESS

Step 1. Identify the Structural & Functional Context
Determine the **key substructures or physicochemical motifs** that likely contributed to the model's decision.
Abstract these into **functional roles**, such as:

- **hydrogen bond donors/acceptors critical for binding to relevant biological targets**,
- **hydrophobic pockets enhancing active site occupancy**,
- **basic or acidic motifs influencing solubility and membrane permeability for intracellular target access**.

> Include both **chemical identity** (what it is) and **functional role** (what it should have done, and how it failed).

Step 2. Evaluate CoT-Truth Misalignment
Compare the model's CoT reasoning with the true label:

- Identify **which reasoning components were correct** (reinforcing aspects);
- Identify **which reasoning components were wrong or misleading** (corrective aspects).

For each incorrect point, analyze **why** it failed – whether due to overestimated hydrophobicity, underestimated desolvation cost, ignored steric clashes in the binding site, or failure to account for critical hydrogen bonding interactions.

Then, summarize how these local reasoning errors led to the global misclassification.

Step 3. Mechanistic Root Cause
Describe the **causal physicochemical and structural mechanism** responsible for the misprediction in {...Dataset-Specific Information}.
You should consider the following critical factors as **guiding hints** (including but not limited to):

- **Target Recognition & Active Site Compatibility**:
(e.g., Did the model fail to recognize hydrophobic cores, polar vectors, or scaffold geometry necessary for productive binding to the relevant biological target?)
- **Conformational Adaptability & Pharmacophore Alignment**:
(e.g., Did the molecule's flexibility, semi-rigid scaffold, or spatial arrangement of functional groups fail to support induced-fit binding?)

Your explanation should clarify:

- **Which feature's contribution the model over- or under-weighted**;
- **Why that led to an incorrect global prediction for {...Dataset-Specific Information}**.

Step 4. Formulate the Decision-Triggering Knowledge Rule

Based on your diagnosis in Step 3, formulate a precise rule that captures this correction using the following universal structure:

- ``principle_id``: (CRITICAL NAMING CONVENTION: Use **'DT_XX_NEW'** for a new rule).
- ``trigger_condition``:
Construct a strictly structured condition string using the exact format: **"[WHO] + [WHERE] + [WITH WHOM]"**.

```

**CRITICAL INDEXING CONSTRAINT (Structural Objectivity):** This field acts as a search index.
  It must describe **"What exists" and **"Where it is", but NEVER **"How it behaves"
  (State/Outcome)**.

**GENERALIZATION REQUIREMENT:** You must abstract the specific atoms into **representative
  chemical classes** to ensure the rule applies to future analogs.

1. **[WHO] Identity:** The functional group class or core motif.
2. **[WHERE] Topology:** The structural anchor.
3. **[WITH WHOM] Environment:** The neighbors.

> **STATE-DESCRIPTION BAN:** You are **STRICTLY PROHIBITED** from using subjective state
  adjectives or outcome-based verbs in this field.

> **NO NARROW EXAMPLES:** Do **NOT** include parenthetical examples in the trigger string.

> Constraint:
> 1. NO NUMERICAL THRESHOLDS.
> 2. QUALITATIVE CONVERSION only.
> 3. The final output string MUST use a **plus sign ('+')** with single spaces.
> 4. Only include chemically necessary factors.

- `principle_statement`:
  A concise mechanistic rationale describing **why** the structural combination (Who+Where+With
  Whom) results in the observed state or outcome.

- `expected_decision_effect`:
  Explicitly state the **Direction of Adjustment** for the contribution weighting and the
  **Target** of that adjustment.

- `adaptive_weighting_guideline`:
  Define the **Magnitude of Adjustment** required (minor attenuation, moderate adjustment,
  major reweighting).

---

## FINAL OUTPUT FORMAT (JSON)
Output only the JSON object.
{
  "justification_of_action": "A brief, 1-2 sentence explanation of your action.",
  "extracted_knowledge": [
    {
      "principle_id": "...",
      "trigger_condition": "[WHO] + [WHERE] + [WITH WHOM]",
      "principle_statement": "...",
      "expected_decision_effect": "...",
      "adaptive_weighting_guideline": "..."
    }
  ]
}

---

## INPUT DATA

## FAILED ATTEMPT DATA

* **Failed CoT Analysis:**
{cot_analysis}

## ORIGINAL MISPREDICTION DATA
* **Mispredicted Molecule**
* SMILES: {smiles}
* Ground Truth Label: {ground_truth_label}
* Model Prediction: {predicted_label}
* **SMILES_description:** {SMILES_description}
* **SMILES_graph_description:** {SMILES_graph_description}
* **SMILES_space_description:** {SMILES_space_description}
* **Molecular_property:** {SMILES_property}
* **valid_sub_smiles:** {valid_sub_smiles}
* **valid_contributions:** {valid_contributions}

```

E.2 Updating CCSAR Repository

We employ this maintenance prompt to integrate new rules into the repository by merging synonymous entries and resolving conflicts, allowing for low-cost knowledge evolution.

Update Rule Base Prompt

You are a medicinal chemistry expert analyzing whether a given molecule acts as a compound with ****{...Dataset-Specific Information}****.

Your goal is to generate objective, mechanism-based reasoning that integrates:

- Global molecular physicochemical properties,
- Substructure composition and their contextual interplay,
- Quantitative contribution values derived from masking-based ML computation.

=====

BACKGROUND (contribution definition)

=====

Each contribution value is computed by masking the substructure, re-predicting the property probability with an ML model, and processing the probability change into a numeric score.

- Positive → the substructure, in this molecule's context, increases predicted {...Dataset-Specific Information} probability.
- Negative → the substructure, in this molecule's context, decreases predicted {...Dataset-Specific Information} probability.
- Magnitude → the absolute value reflects effect strength; values near zero indicate weak effect.
- These contributions are ****context-dependent**** (they are computed within the full molecular environment and implicitly reflect fragment interactions).
- The provided list may not cover all fragments in the molecule; treat it as a focused, partial attribution map.

IMPORTANT: **Do not decide activity by naïvely summing these numbers.** The purpose of these values is to ***point*** the model to which fragments to examine and to suggest relative influence; you must explain ***why*** each fragment's value is positive/negative and why its magnitude is as reported, in the context of the whole molecule.

=====

TASK INSTRUCTIONS

=====

You must perform structured molecular reasoning before answering.

Note: A Reference_Label is provided for orientation only to help the generator learn correct-style outputs. ****Do not mention, quote, or reveal the Reference_Label anywhere in your reasoning.**** The model's final decision must be justified solely by its integrated reasoning and may differ from Reference_Label.

For each substructure in VALID_SUB_SMILES (briefly, 1-2 short clauses each):

- State the reported contribution value and its sign.
- Explain, using evidence from SMILES, GRAPH_DESCRIPTION, SPACE_INFORMATION, and PROPERTY_DESCRIPTION, ****why**** this fragment shows that sign in this molecular context (e.g., hydrophobic aromatic systems promoting binding into relevant hydrophobic pockets of the biological target, basic amines forming salt-bridges with target active sites, polar/ionized groups reducing cell uptake, metal-chelating moieties interfering with the target mechanism, bulky polar substituents hindering target binding or cell permeation).
- Explain ****why**** the reported magnitude is reasonable (e.g., a large hydrophobic scaffold that complements a target pocket → large positive; a strongly ionized polar headgroup that prevents cell entry → strong negative).

Then:

- Provide a focused analysis of the ****global molecular properties**** in PROPERTY_DESCRIPTION, ****but only comment on those properties that have a materially important effect**** on potency and cell-based activity for this molecule (e.g., LogP if it affects cellular membrane permeability and intracellular exposure, TPSA if it limits cell uptake, MW if it restricts diffusion or bioavailability, pKa if it alters ionization and target interaction). For each such selected property, state how that specific value mechanistically supports or opposes {...Dataset-Specific Information}. Do not list or comment on properties that are irrelevant or have negligible impact.

- Describe how these fragment-level effects **interact** (synergy, cancellation, internal H-bonding masking polarity, conformational effects enabling target fit) and how global properties (LogP, TPSA, MW, HBD/HBA, rotatable bonds, pKa) modify or validate those fragment-level interpretations.
- Do NOT invent arbitrary numeric thresholds or external data.

Reasoning constraints:

- Inside `<think>` be concise but explicit: each substructure explanation should be one short clause; the interaction/global synthesis may be 2-4 short sentences. Total `<think>` length: 120-220 words.
- Do NOT use hedging/speculative words ("maybe", "could", "might", "seems", "likely", "possibly", "perhaps", "uncertain").
- Do NOT perform a raw arithmetic sum of contribution values to decide activity.

NEW OUTPUT RULE (required):

- The `<think>` section **must end** with a single-line explicit decision in the format: `<Decision: yes>` or `<Decision: no>`. This decision line counts as part of `<think>` and must reflect the integrated reasoning.

```
=====
### OUTPUT FORMAT (strict)
=====
```

```
<think>
Per-substructure lines: "Fragment X (SMILES_or_label): contribution_value → brief explanation
of sign and magnitude."
Selective property analysis lines: "Property Y (value): brief mechanistic interpretation
supporting or opposing {...Dataset-Specific Information} (only include properties with
material impact)."
Synthesis: 2-4 short sentences integrating interactions and global properties.
Decision line: "Decision: yes" or "Decision: no"
</think>
```

```
<answer>yes</answer> or <answer>no</answer>
```

Critical rules for `<answer>`:

- `<answer>` must be exactly one token: `<yes>` or `<no>` (lowercase), matching the `<Decision:>` line in `<think>` exactly.
- No additional text, punctuation, or commentary in `<answer>`.

FORBIDDEN:

- No text outside the two tags.
- No markdown, headings, notes, or external assumptions.

```
=====
### INPUT FIELDS
=====
```

```
SMILES: {smiles}
SMILES_DESCRIPTION:
{smiles_description}
GRAPH_DESCRIPTION:
{graph_description}
SPACE_INFORMATION:
{space_description}
PROPERTY_DESCRIPTION:
{property}
VALID_SUB_SMILES: {valid_sub_smiles} # A list of substructures related to
{...Dataset-Specific Information}.
VALID_CONTRIBUTIONS: {valid_contributions} # Numerical contribution values aligned with
VALID_SUB_SMILES.
Reference_Label: {label} # Provided for orientation only – DO NOT mention
or quote this label in the reasoning. The final answer must be based on the model's own
integrated reasoning (it may agree or disagree with Reference_Label).
```

F Prompt Templates for Constructing CCRT

F.1 Generating 3W Descriptions from Molecular Structures

These templates are used to generate comprehensive "3W" (What, Why, and How) descriptions that capture the underlying chemical logic and structural influences of a molecule.

Get Anchor Prompt

```
# ROLE: Medicinal Chemistry Knowledge Engineer ({...Dataset-Specific Information})

# TASK
Translate specific molecular data into Generalized Structural Queries that map to an
{...Dataset-Specific Information} Rule Base.

# OBJECTIVE
The Rule Base matches generalized structural and chemical classes based on
{...Dataset-Specific Information} mechanisms.
You must generate queries that describe the structural and chemical motifs driving these
mechanisms:

CORE PHILOSOPHY (Structure-State Logic):
* Step 1: Check Presence: Does the molecule contain the specific substructure class
  relevant to the assay's biological mechanism?
* Step 2: Describe State: If the substructure exists, you MUST generate a query
  describing its actual state (e.g., is the lipophilic core planar or twisted? is the
  electrophile steric-shielded or exposed?). Do not skip just because it looks "safe".
* Step 3: Skip if Absent: Only skip a dimension if the molecule completely lacks the
  relevant substructure class.

Note: The specific scenarios, biological targets, or mechanisms mentioned in any examples
below are illustrative demonstrations only and are NOT exhaustive. Do not limit your
analysis to these specific instances. You must derive the required structural state based on
the actual chemical and physical reality of the specific molecule provided.

### Dimension 1: {...Dataset-Specific Information}
* Critical Mechanism: {...Dataset-Specific Information}
* Analysis Focus: {...Dataset-Specific Information}

### Dimension 2: {...Dataset-Specific Information}
* Critical Mechanism: {...Dataset-Specific Information}
* Analysis Focus: {...Dataset-Specific Information}

### Dimension 3: {...Dataset-Specific Information}
* Critical Mechanism: {...Dataset-Specific Information}
* Analysis Focus: {...Dataset-Specific Information}

---

# MANDATORY GENERALIZATION PROTOCOL (The "Class" Filter)
You MUST convert specific atoms/groups into their Functional Class to ensure
generalization.

* FORBIDDEN: Specific atom/group names
  (e.g., "Chlorine", "Methyl", "Phenyl", "Piperidine", "Pyridine").

* REQUIRED: Functional descriptions

---

# THE "3W" QUERY FORMAT
Structure every query strictly as: [WHO] + [WHERE] + [WITH WHOM]

* [WHO] Identity: The functional group class, pharmacophore or core motif (e.g., "Polar
  chelating motif" or "Basic heterocycle").
* [WHERE] Topology: The structural anchor (e.g., "on aryl scaffold").
* [WITH WHOM] Environment: The neighbors (e.g., "adjacent to bulky lipophilic
  substituent").

Each query must contain exactly two "+" signs.

---

# REFLECTION & SAFETY CHECK
Before outputting, verify:

1. Generalization: All atom names replaced by functional classes.
```

```

2. 3W Structure: Each query contains exactly two + signs and follows [WHO] + [WHERE] + [WITH WHOM].
3. Mechanism Coverage: Queries collectively reflect the four HIV-specific chemical mechanisms/constraints defined above.
4. Specificity Requirement: Each generated query must be grounded in the given molecule's specific features as they relate to the critical mechanisms.
5. Single-Concept Requirement: Each generated query should express a single core concept; avoid redundancy of information across queries.

---

# OUTPUT FORMAT
Return ONLY a JSON object with a single key `queries`, containing a list of the generalized query strings. (less than 5 queries for each molecular)

Example Output Format:
{{
  "queries": [
    "...",
    "...",
    "...",
    "...",
    "..."
  ]
}}

---

# INPUT DATA
* 1. SMILES: {smiles}
* 2. Molecular Text Description: {molecular_description}
* 3. Graph Topology: {graph_desc}
* 4. Spatial Conformation: {space_desc}
* 5. Physicochemical Properties: {property_desc}

```

F.2 Constructing CCRT

The final template for CCRT construction instructs the model to synthesize retrieved rules with molecular descriptions, ensuring that the reasoning paths are specifically calibrated to the target molecule's unique environment.

Refine CoT Prompt

```

# ROLE: Expert Mechanistic Analyst ({...Dataset-Specific Information})

## YOUR TASK
You are a senior medicinal chemist and DMPK expert acting as a CoT optimizer for the {...Dataset-Specific Information} prediction task.
Your goal is to refine an existing reasoning chain (Baseline CoT) by integrating mechanism-based and structure-property knowledge relevant to {...Dataset-Specific Information}, while keeping the overall reasoning format, expression style, and logical flow consistent with the original CoT.

This optimized CoT will serve as high-quality mechanistic reasoning data for training molecular reasoning models.

---

## OBJECTIVES

1. Minimal Deviation: Keep the phrasing, tone, and logical progression close to the Baseline CoT.
   The optimization should feel like a natural continuation or correction from the same analyst.

2. Mechanistic Integration: If a retrieved knowledge rule is triggered by a specific fragment, append two additional fields for that fragment:
   - `Mechanistic Refinement`

```

- `Final Rationale`

These fields must clearly show *how* the knowledge modifies or supports the baseline interpretation regarding cellular exposure, efflux/uptake, target binding (reverse transcriptase, protease, integrase, entry), or metabolic stability.

If no rule is triggered, keep only the original reasoning structure – but still perform subtle stylistic and logical improvements based on your expertise (clarity, precision, or minor adjustments).

3. **Consistency with Training Logic:** The optimized CoT must preserve the original fragment-based reasoning and contribution-driven explanation.
4. **Independent Judgment:** The final decision must follow naturally from your reasoning.
5. **Intelligent Trigger Scrutiny (Guardrail for Trigger Hallucination):**

This is your most important reasoning task for **Step 1**. Before triggering any rule, you must first determine its **intent**: is it a **'Specific Fact'** or a **'General Principle'**?

 - * **A. 'Specific Fact' Rules (Strict Matching):**
 - * **Definition:** Rules that name a *specific, unambiguous* chemical structure as a hard requirement (e.g., "chlorinated ring", "pyridine", "gem-difluoro-oxazoline").
 - * **Action:** These **must** be evaluated with **absolute, literal, atom-for-atom strictness**. A "methyl" group **CANNOT** trigger a "chlorinated" rule. If the literal atoms do not match, the rule **must** be considered "Not Triggered" and **must not** be mentioned in the `Knowledge Evaluation` field.
 - * **B. 'General Principle' Rules (Generalization Matching):**
 - * **Definition:** Rules that describe a *mechanism, process, or pattern* (e.g., "shielding by flexible chains", "high desolvation penalty", "efflux recognition motif", "metal-chelation enabling integrase binding"). These rules often use examples like "*e.g.*", "*such as*", or "*or*".
 - * **Action:** These **SHOULD** be triggered by *analogous* structures that fit the *principle*. This is **desirable generalization**. (e.g., The *principle* "shielding of a polar pharmacophore by flexible chains" correctly triggers for a "sulfone core" even if the *example* given was an "amide core".)
6. **Selective Application (Guardrail for Application Hallucination):**

This is your most important reasoning task for **Step 2**. **After** a rule is correctly triggered (per Objective 5), your reasoning **must** be specific to the molecule.

 - * **Action:** When writing the `Mechanistic Refinement` or `Final Rationale`, you **must** only reference the structural features that **actually exist** in the molecule.
 - * **Strict Prohibition:** **Do not** copy, cite, or hallucinate any part of the rule's *example text* (e.g., the word "amide") if that structure is not present in the molecule. Your reasoning must be grounded in the **molecule**, not in the rule's general text.

REFINEMENT INSTRUCTIONS

Step 1: Knowledge Triage (Knowledge Evaluation)

- * **Your first action** inside the ``<think>`` block is to perform a triage of all `Retrieved Knowledge`.
- * **If one or more rules ARE triggered:**
 - * Create the `Knowledge Evaluation` field.
 - * For *each triggered rule*, list the rule, the structural feature that triggered it, and the brief implication for antiviral potency or intracellular exposure.
- * **If NO rules are triggered:**
 - * Create the `Knowledge Evaluation` field.
 - * State this explicitly.

Step 2: Fragment-by-Fragment Refinement

For each fragment mentioned in the Baseline CoT, you **must strictly** choose **ONE** of the following two paths:

PATH A: If a rule was triggered for this fragment (as per Step 1):

1. Output the `Fragment` line with its refined baseline reasoning.

2. ****Then****, output the `Mechanistic Refinement` field, explaining the rule's impact on antiviral activity or cellular exposure (e.g., changes to desolvation cost, efflux liability, intracellular accumulation, target engagement geometry).
3. ****Then****, output the `Final Rationale` field, synthesizing the net effect.

****PATH B: If NO rule was triggered for this fragment:****

1. Output ****ONLY**** the `Fragment` line, containing just the refined baseline reasoning (e.g., "Fragment *[name]*...: [refined logic, e.g., 'contribution to lipophilicity helps cell membrane traversal and intracellular exposure']").
2. You ****MUST NOT**** output the `Mechanistic Refinement` or `Final Rationale` fields for this fragment. Do not even output them with "no rule triggered" inside; the fields themselves must be absent.

Step 3: Synthesis and Mechanistic Trade-off (MANDATORY)

This step ****must**** perform the explicit competitive analysis, following your defined hierarchy ("heuristic as primary, rules as supplemental"). You will output a ****single field**** named `Synthesis and Mechanistic Trade-off` that contains this entire logic.

1. ****Summarize Heuristic Principle:**** In your mind, first assess the **Heuristic Evidence**. This is the **principle-based description** from the fragment scores.
2. ****Summarize Knowledge-Based Adjustments:**** In your mind, identify the **qualitative** adjustments from the triggered Knowledge Rules.
3. ****Perform Trade-off (Heuristic vs. Knowledge):**** This is the crucial step. You must now explicitly weigh the **Heuristic Principle (1)** against the **Knowledge-Based Adjustments (2)**. ****This adjustment CAN BE SUBVERSIVE.**** A Knowledge Rule (2) ****can and should**** **override** or **invert** a Heuristic Principle (1) if the mechanistic evidence is strong enough (e.g., high lipophilicity but specific P-gp efflux substrate or strong metabolic liability).
4. ****Write the `Synthesis and Mechanistic Trade-off` field:**** This field ****must**** be written to contain this full analysis, following this specific order:
 - ***(a) Heuristic Conclusion:**** Start by stating the conclusion suggested by the quantitative scores AND their underlying **principle-based description**.
 - ***(b) Knowledge Adjustments:**** State the penalties (Cons) and compensations (Pros) introduced by the retrieved knowledge rules.
 - ***(c) Final Trade-off:**** Explicitly state the trade-off **between** the heuristic (a) and the rules (b).

Step 4: Final Verdict and Decision

This step now **follows** the trade-off and must be consistent with it.

1. ****Final Verdict Justification (Mandatory):****
 - *** Create the `Final Verdict Justification` field.
 - *** This field ****must**** be a concise, one- or two-sentence summary that is ****100% consistent**** with the "Trade-off" verdict (part (c)) made in `Step 3`.
2. ****Logical Coherence:****
 - *** Ensure the final decision ("yes" for Active / "no" for Inactive against HIV) (in the next field) follows **organically** from the `Synthesis and Mechanistic Trade-off` and `Final Verdict Justification`.
3. ****Decision (Mandatory in ``<think>`` block):****
 - *** Create the `Decision` field inside the ``<think>`` block.
 - *** This field ****must**** contain the `yes` or `no` that matches the `Final Verdict Justification`.

OUTPUT FORMAT

Return the optimized reasoning in **exactly this format**:

`<think>`

****Knowledge Evaluation****: [If rules were triggered, list them and their justification here. If NO rules were triggered, state that explicitly.]**

[FOR A FRAGMENT ****WITH**** A TRIGGERED RULE (Follow Step 2, Path A):]

Fragment *[fragment1]*...: [refined baseline reasoning (the principle-based description)].

Mechanistic Refinement: [Applying Rule DT_XXX, describe mechanistic insight relevant to antiviral potency or intracellular exposure].

Final Rationale: [How the rule modifies this fragment's role].

[FOR A FRAGMENT ****WITHOUT**** A TRIGGERED RULE (Follow Step 2, Path B):]
Fragment `*[fragment2]...*`: [Output ****ONLY**** the `Fragment` line, containing just the refined baseline reasoning.]

(...repeat for all fragments, strictly choosing the correct format for each...)

Property [name] ([value]): [refined reasoning sentence].
(...other properties...)

****Synthesis and Mechanistic Trade-off****: ****a) Heuristic Principle****: [Start by stating the conclusion from the quantitative scores AND their underlying **principle**.] ****b) Knowledge Adjustments****: [State the 'Pros' (compensations) and 'Cons' (penalties) introduced by the retrieved Knowledge Rules.] ****c) Trade-off****: [State the final verdict **between** the heuristic (a) and the rules (b), and **why** one side wins. Explicitly state if the adjustment is **reinforcing** or **subversive**, and if the flaw is **(non-)compensable**.]

****Final Verdict Justification****: [Must be a concise, 1-2 sentence summary of the "Trade-off" (part (c)) from the Synthesis. e.g., "The balance favors intracellular activity because heuristic lipophilicity evidence was reinforced by rules reducing efflux liability."]

Decision: [yes/no]
</think>
<answer>[yes/no]</answer>

INPUT DATA

* ****Retrieved Knowledge (Rules to be Evaluated)****:
* extracted_knowledge:
{extracted_knowledge}
* ****Purpose****: A list of **potential** rules. You must evaluate **if** they are triggered before applying them.

* ****Molecule Context****:
* SMILES: {smiles}
* SMILES_description: {SMILES_description}
* SMILES_graph_description: {SMILES_graph_description}
* SMILES_space_description: {SMILES_space_description}
* SMILES_property_description: {property_description}
* VALID_SUB_SMILES: {valid_sub_smiles}
* VALID_CONTRIBUTIONS: {valid_contributions}
* Ground Truth Label: {label}
* ****Purpose: For reference and context ONLY.****

* ****Baseline CoT (Initial Draft)****:
* cot_original:
{cot_original}
""