MolRAG: Unlocking the Power of LLMs for Molecular Property Prediction

Ziting Xian
Sun Yat-sen University
`xianzt@mail2.sysu.edu.cn`

Jiawei Gu
Sun Yat-sen University
`kuvvius@gmail.com`

Lingbo Li
University of Warwick
`lingbo.li.1@warwick.ac.uk`

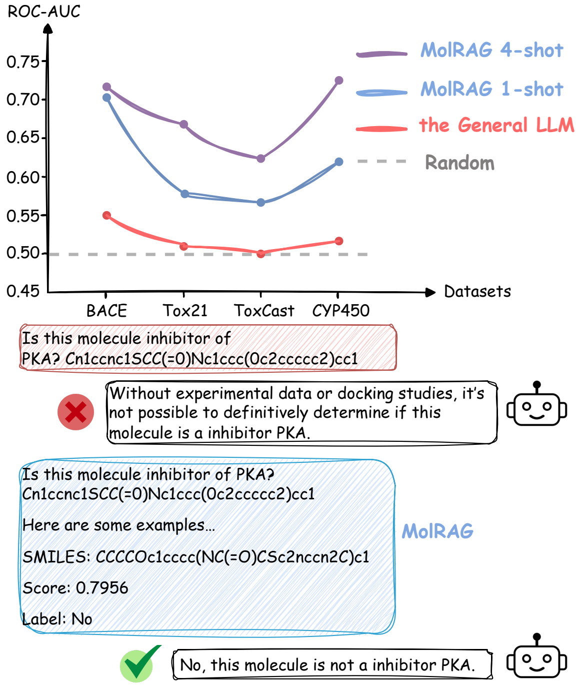
Shangsong Liang[*]
Sun Yat-sen University
`liangshangsong@gmail.com`

Abstract

Recent LLMs exhibit limited effectiveness on molecular property prediction task due to the semantic gap between molecular representations and natural language, as well as the lack of domain-specific knowledge. To address these challenges, we propose MolRAG, a Retrieval-Augmented Generation framework integrating Chain-of-Thought reasoning for molecular property prediction. MolRAG operates by retrieving structurally analogous molecules as contextual references to guide stepwise knowledge reasoning through chemical structure-property relationships. This dual mechanism synergizes molecular similarity analysis with structured inference, while generating human-interpretable rationales grounded in domain knowledge. Experimental results show MolRAG outperforms pre-trained LLMs on four datasets, and even matches supervised methods, achieving performance gains of 1.1%–45.7% over direct prediction approaches, demonstrating versatile effectiveness. Our code is available at https://github.com/AcaciaSin/MolRAG.

1 Introduction

Molecular property prediction (Wu et al., 2018) is a fundamental task in computational chemistry and drug discovery, aiming to predict the quantitative characteristics of chemical compounds. Accurate prediction of molecular properties enables researchers to screen potential drug candidates at early stages, substantially reducing experimental costs and enhancing drug discovery efficiency. Current deep learning-based molecular property prediction models face two fundamental challenges (Kipf and Welling, 2022; Veličković et al., 2018; Xu et al., 2018): (1) strong dependency on scarce annotated data constrains their applicability in low-resource scenarios; (2) insufficient explicit modeling of physicochemical principles compromises the interpretability of prediction outcomes.



Is this molecule inhibitor of PKA? Cn1ccnc1SCC(=O)Nc1ccc(Oc2ccccc2)cc1

✗ Without experimental data or docking studies, it's not possible to definitively determine if this molecule is a inhibitor PKA.

Is this molecule inhibitor of PKA? Cn1ccnc1SCC(=O)Nc1ccc(Oc2ccccc2)cc1

Here are some examples...

SMILES: CCCCOc1cccc(NC(=O)CSc2nccn2C)c1

Score: 0.7956

Label: No

✓ No, this molecule is not a inhibitor PKA.

Figure 1: Performance comparison of MolRAG and General LLMs. The blue and purple lines in the line chart show the results of the MolRAG 1-shot and 4-shot, and the red line shows the results of the General LLM.

To overcome the challenges, recent researchers have begun exploring the potential of large language models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Yang et al., 2024; Gu and Liang, 2025) for molecular property prediction. LLMs address the molecular property prediction task by encoding structural information (e.g., SMILES strings, molecular graphs) into training data (Zeng et al., 2022; Su et al., 2022; Taylor et al., 2022; Zhao et al., 2023). However, existing LLMs exhibit notable limitations: First, the semantic gap between hierarchical molecular representations and natural language hinders the effective capture of critical chemical features such as functional groups. Second, inadequate implicit encoding of domain-specific knowledge (e.g., physicochemical laws) restricts the reasoning reliability of the general LLMs. As illustrated in Figure 1, the general LLMs per-

---

[*] Shangsong Liang is the corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15513–15531
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

form the molecular property prediction task poorly due to the two limitations.

To overcome these limitations, we propose **MolRAG**, a Retrieval-Augmented Generation (RAG) LLM framework combined with Chain-of-Thought (CoT) reasoning to enhance molecular property prediction. This approach starts with dynamically retrieving analogous molecules from specified knowledge bases, and then integrates them into the context for step-by-step reasoning. Instead of relying solely on internal knowledge, this retrieval mechanism strengthens domain knowledge integration, improves predictive accuracy, and thereby enhances adaptability across different molecular property tasks and datasets. Unlike conventional pre-training or fine-tuning methods, MolRAG requires no additional training, which significantly reduces computational costs. Experimental results show that MolRAG outperforms pre-trained LLMs on 4 general datasets, and on 2 of them even surpasses existing GNNs-based methods with higher prediction accuracy.

In summary, our contributions can be summarized as follows: (1) **Training-Free Molecular Property Prediction.** MolRAG enables LLMs to perform molecular property prediction without the need for large-scale pre-training or fine-tuning, significantly reducing computational costs. (2) **Interpretable and Adaptive Reasoning.** By incorporating CoT reasoning and retrieved molecular knowledge, MolRAG enhances predictive accuracy while providing interpretable and insightful explanations. (3) **Robust Retrieval-Augmented Framework.** Through systematic evaluation, we demonstrate the effectiveness of different molecular retrieval strategies and in-context learning in improving LLM reasoning for chemical tasks.

2 Related Work

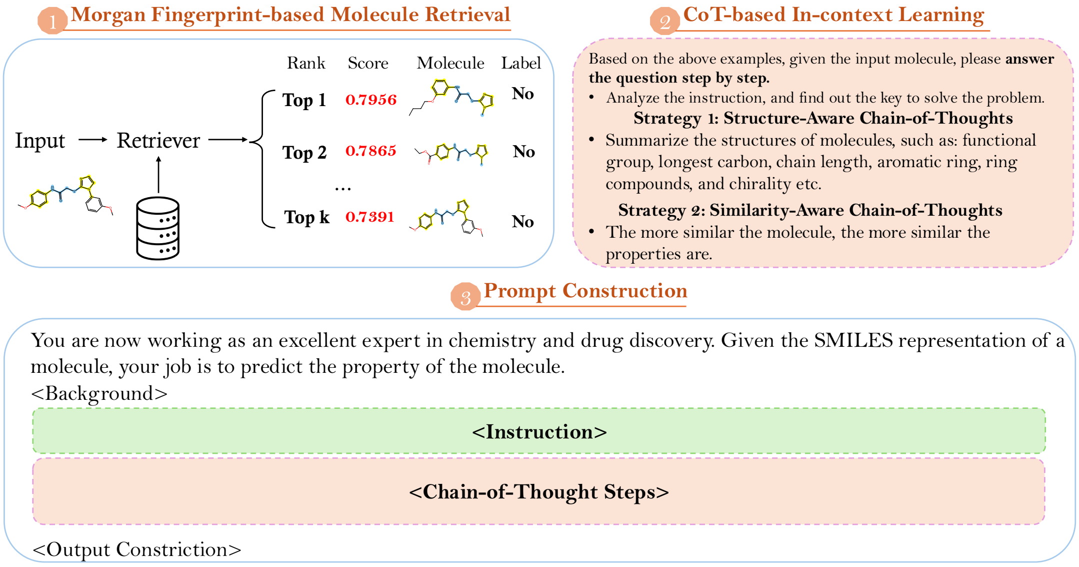
LLMs for molecular property prediction. Despite the strong capabilities of LLMs, directly applying LLMs to molecular property prediction task still face challenges (White et al., 2023; Castro Nascimento and Pimentel, 2023; Guo et al., 2023). To overcome these challenges, earlier studies often employed specialist language models pre-trained from scratch (Zeng et al., 2022; Su et al., 2022; Taylor et al., 2022; Zhao et al., 2023), or through instruction tuning on specific tasks (Fang et al., 2023; Cao et al., 2023). Our work focuses on how to use a training-free approach, leveraging existing LLMs to complete molecular property prediction tasks. By exploring training-free methods, we aim to fully harness the generality of LLMs while reducing the dependency on specialized data and computational resources.

Retrieval augmented generation in chemistry. LLMs face challenges when dealing with specialized or knowledge-intensive tasks. In these tasks, LLMs are often prone to producing "hallucinations (Huang et al., 2023)". Retrieval-augmented generation enhances LLMs by retrieving relevant content from external knowledge databases (Lewis et al., 2020). Recently, retrieval-augmented generation methods have also been applied to chemistry tasks. Li et al. (2024a) utilized retrieval to complete molecular captioning and molecular generation tasks. ChatDrug (Liu et al., 2024b) applied LLMs for drug editing tasks. ChemCrow (M. Bran et al., 2024) integrated various retrieval tools to finish general chemical tasks. Despite these, no approach has been developed to target the performance of LLMs in molecular property prediction tasks by leveraging the RAG framework.

CoT reasoning for chemistry. Chain-of-thought (CoT) aims to guide the model through a series of intermediate reasoning steps to obtain the final answer (Wei et al., 2022; Kojima et al., 2022). A few works have also applied CoT reasoning in the chemistry domain for protein interactions (Jin et al., 2024) and chemistry question-answering (Ouyang et al., 2024). However, none of these works use chain-of-thought reasoning in molecular property prediction tasks. MolRAG focuses on molecular property prediction tasks, not only applying retrieval-augmented methods to molecular property prediction but also innovating on how to integrate retrieval information to formulate a CoT reasoning strategy. This approach enhances the reasoning ability and performance of LLMs in molecular property prediction tasks.

3 Our MolRAG Model

To fully leverage the capabilities of LLMs and incorporate the principle of molecular structure-property relationship, we propose MolRAG, a training-free, retrieval-augmented generation framework. The overall framework is depicted in Figure 2.



Based on the above examples, given the input molecule, please **answer the question step by step.**

- Analyze the instruction, and find out the key to solve the problem.

Strategy 1: Structure-Aware Chain-of-Thoughts

- Summarize the structures of molecules, such as: functional group, longest carbon, chain length, aromatic ring, ring compounds, and chirality etc.

Strategy 2: Similarity-Aware Chain-of-Thoughts

- The more similar the molecule, the more similar the properties are.

You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule.
<Background>
<Instruction>
<Chain-of-Thought Steps>
<Output Constriction>

Figure 2: Overall Framework of MolRAG. MolRAG begins with Morgan Fingerprint-based Molecule Retrieval to obtain example molecules, followed by CoT-based In-context Learning for CoT strategy selection, and ends with Prompt Construction to formulate the prompt.

3.1 Problem and Overview of MolRAG

Molecular property prediction aims at giving an input molecule I, predicting the properties of the molecule. In MolRAG, the retriever R first retrieves relevant contents from the database D based on the input molecule I by computing the similarity between I and molecules in the database. The retrieved molecules are ranked by similarity scores in descending order, and the top k most similar molecules $K = R(I|D)$ are identified, where k ranges from 0 to n. The corresponding similarity scores are recorded as $Score\ i$, and the labels are recorded as $Label\ i$.

Next, MolRAG uses the CoT-based in-context learning strategy S to further complete the prompts. There are two strategies in MolRAG, which are Structure-Aware Chain-of-Thoughts, denoted as $Struct\text{-}CoT$, and Similarity-Aware Chain-of-Thoughts, denoted as $Sim\text{-}CoT$.

Finally, MolRAG constructs the few-shot prompt using a structured template $Prompt = (Instruction, I, K, S)$. This prompt is sent to the LLM to output the property of I.

3.2 Morgan Fingerprint-based Retrieval

The Molecular Fingerprint (Butina, 1999) is widely used for molecular similarity searches because it explicitly encodes key chemical features, allowing for fast comparison and similarity calculations across large databases.

The Morgan Fingerprint is a specific type of molecular fingerprint that is commonly used (Zhou and Skolnick, 2024). It is generated through molecular graph traversal and is particularly effective in capturing local substructure features of molecules. Specifically, the Morgan Fingerprint creates a feature vector by recursively recording the connectivity information of each atom and its neighboring atoms within the molecule. The resulting fingerprint is represented as a fixed-length binary or integer vector, which efficiently encodes both the global and local structural features of the molecule

The molecular similarity is quantified using the Dice coefficient, which measures feature overlap between fingerprint vectors:

$$\text{Dice Similarity}(A, B) = \frac{2|A \cap B|}{|A| + |B|}, \qquad (1)$$

where A and B represent a molecule from the database and the current input molecule, respectively. $A \cap B$ is their intersection, and $|A|$ and $|B|$ are their sizes (i.e., the number of elements). The Dice Similarity range from 0 to 1, when the value is closer to 1 indicating a higher structural similarity between the two molecules.

In summary, the molecular retrieval in MolRAG is as follows: (1) **Fingerprint Generation:** For each molecule, the Morgan algorithm generates the corresponding molecular fingerprint. (2) **Similarity Calculation:** The Dice Similarity is used to calculate the similarity between the target molecule and each molecule in the database. (3) **Ranking and Selection:** Based on the calculated similarity scores, the molecules are ranked, and the top K most similar molecules are selected as candidate molecules.

3.3 CoT-based In-context Learning

Chain-of-Thought (CoT) improves the reasoning ability of LLMs in molecular property prediction by decomposing complex SMILES data processing into stepwise sub-tasks, overcoming the limitations of direct question-answering approaches. Furthermore, MolRAG introduces a CoT-based prompt construction method, marking the first time the CoT is applied to molecular property prediction tasks. This method guides the LLM to answer questions step by step, gradually reasoning through the solution, especially when combined with the retrieved molecular data. This step-by-step reasoning strategy enables the LLM to better understand the complex task requirements, ultimately resulting in more accurate prediction.

Moreover, the CoT-based prompt construction method improves reasoning accuracy and enhances the interpretability of the generated content. By explicitly stating the rationale for each reasoning step, researchers can gain a clearer understanding of the model's decision-making process. Interpretability is especially important in molecular property prediction tasks, as it helps reveal the underlying mechanisms driving the predictions and provides valuable insights for further optimization and informed decision-making.

To further explore how to perform CoT-based in-context learning using the retrieved content, MolRAG has adopted two different strategies for constructing in-context learning.

3.3.1 Structure-Aware Chain-of-Thoughts

Structure-Aware Chain-of-Thought (Struct-CoT) establishes a reasoning framework that connects molecular structural patterns to their macroscopic properties. By explicitly incorporating chemical structure-property relationships into the reasoning pathway, this approach allows for a systematic interpretation of how specific structural features (e.g., functional groups, stereochemistry) influence molecular properties. This significantly enhances prediction interpretability, providing a clearer understanding of molecular behavior compared to black-box deep learning methods.

In MolRAG, we implement Struct-CoT through hierarchical feature-guided reasoning:

- **Guided Structural Decomposition:** The LLM systematically extracts critical structural features (e.g., functional groups, aromatic ring and chirality) from both target and retrieved molecules.
- **Property Correlation Mapping:** These features are explicitly linked to target properties through physicochemical principles. For instance, aromatic rings enhance hydrophobicity and thermal stability, while strained cyclic structures (e.g., norbornene) exhibit elevated chemical reactivity.
- **Comparative Reasoning:** Structural similarities and differences between molecules are leveraged to deduce variations in properties.

Struct-CoT ensures that the model's predictions are grounded in verifiable chemical knowledge rather than statistical correlations.

3.3.2 Similarity-Aware Chain-of-Thoughts

Similarity-Aware Chain-of-Thought (Sim-CoT) establishes a retrieval-enhanced reasoning paradigm by explicitly linking molecular structural similarity with property correlations. This framework operationalizes the fundamental chemical principle that "structurally similar molecules exhibit property continuity"(Johnson et al., 1990), guiding the model to systematically leverage retrieved reference molecular information.

In MolRAG, Sim-CoT is implemented through two key mechanisms: (1) **Cross-molecule Correlation:** The LLM is constrained to compare critical structural features between target and retrieved molecules of high similarity scores, analyzing how these features collectively influence target properties. (2) **Property Continuity Constraints:** Interpretable reasoning chains are constructed based on established chemical rules, such as “increasing methyl substitutions on benzene rings progressively enhance hydrophobicity,” to map structural variations to property changes. Sim-CoT facilitates the integration of retrieved molecules into the decision-making process, reducing reliance on mere statistical correlations while ensuring that predictions remain chemically plausible.

3.4 Prompt Construction

Finally, building upon the original instructions, the retrieved information is combined with the selected CoT strategy to generate the final prompt for the

Table 1: Experimental Results on Test Dataset (Classification Tasks). The teal color indicates the best performance of pre-training methods.

Model	Shot	Method	BACE (152)	HIV (4113)	MUV (25342)	Tox21 (7069)	ToxCast (137215)	BBBP (204)	CYP450 (5669)
Llama3-8b	0-shot	/	0.5186	0.5801	0.4502	0.5378	0.4992	0.5226	0.5107
		Struct-CoT	0.5044	0.5136	0.5387	0.5341	0.5016	0.5462	0.5179
	1-shot	Struct-CoT	0.5939	0.5432	0.4962	0.5449	0.5338	0.5457	0.5488
		Sim-CoT	0.7188	0.5749	0.5540	0.5880	0.5683	0.5469	0.6244
	2-shot	Struct-CoT	0.6151	0.5684	0.4963	0.5495	0.5557	0.5642	0.5533
		Sim-CoT	**0.7541**	0.5830	0.5559	0.6279	0.6012	0.5116	0.6903
	4-shot	Struct-CoT	0.6257	0.5948	0.5527	0.5662	0.5913	**0.5723**	0.5838
		Sim-CoT	0.7225	**0.6436**	**0.5616**	**0.6393**	**0.6408**	0.5411	**0.7229**
Random-Choice			0.5409	0.4897	0.5274	0.4989	0.4969	0.5017	0.4964
Pre-training Methods		Gimlet	0.6957	0.6624	0.6439	0.6119	0.5904	0.5939	0.7125
		KVPLM	0.5126	0.6120	0.6172	0.4917	0.5096	0.6020	0.5922
		MoMu	0.6656	0.5026	0.6051	0.5757	0.5238	0.4981	0.5798
		Galactica-125M	0.4451	0.3671	0.4986	0.4964	0.5106	0.6052	0.5369
		Galactica-1.3B	0.5648	0.3385	0.5715	0.4946	0.5123	0.5394	0.4686
Graph-based Networks		GCN	0.7360	0.7570	0.7320	0.7490	0.6330	0.6490	0.8041
		GAT	0.6970	0.7290	0.6660	0.7540	0.6460	0.6650	0.8281
		GIN	0.7010	0.7530	0.7180	0.7400	0.6340	0.6580	0.8205
		Graphormer	0.7760	0.7452	0.7061	0.7589	0.6470	0.7015	0.8436
		Graphormer-p	0.8575	0.7788	0.7480	0.7729	0.6649	0.7163	0.8877

LLM, and then extract the final answer from the LLM output to get the final answer. Although these two CoT strategies differ in the construction of the steps, their core goal is the same: optimizing the model's in-context learning mechanism to enhance its ability to integrate information during reasoning. Within the MolRAG framework, LLM can accurately utilize external knowledge within the given context to gradually deduce molecular properties, thereby improving the reliability and precision of the predictions and providing a more efficient, transparent, and interpretable solution for molecular property prediction tasks.

4 Experiment

In this section, we aim to answer the following research questions: **(RQ1)** What are the advantages of MolRAG compared with the General LLM? **(RQ2)** What's core information contributes to molecular property prediction tasks?

4.1 Experimental Setting

Datasets and Database Construction. We validated our MolRAG on ten datasets, which are from MoleculeNet (Wu et al., 2018), a widely used benchmark for molecular property prediction tasks. These datasets provide a comprehensive evaluation of different aspects of molecular property prediction. The tasks in these datasets are categorized into four types: Physico-chemical tasks, Bio-activity tasks, Toxicity tasks and Pharmacokinetic tasks. For instructions construction process, we adopted instructions from Gimlet (Zhao et al., 2023) and followed the same experimental settings used in the evaluation process. Details of dataset and instructions are in Appendix A. We constructed the database using the Gimlet (Zhao et al., 2023)'s training split of downstream tasks, and using the test split for evaluation. Additionally, considering that some datasets contain an excessive number of instructions, we constructed a test-mini dataset to reduce inference costs. For datasets with more than 4,000 instructions, we randomly sampled 4,000 instructions to form the **Test-Mini** dataset.

Baselines. We verify the performance enhancement of MolRAG in the setting of generalist models. However, considering experimental duration and computational cost, we selected Llama3-8B-Instruct (Touvron et al., 2023) as our primary model to conduct experiments on the **Test** dataset. Meanwhile, GPT-4o (Achiam et al., 2023) and Qwen2.5-7B-Instruct (Yang et al., 2024) are evaluated on the **Test-Mini dataset**. Additionally, we include the pre-training baselines including Gimlet (Zhao et al., 2023), KVPLM (Zeng et al., 2022), MoMu (Su et al., 2022), Galactica-125M (Taylor et al., 2022) and Galactica-1.3B , and supervised baselines including: GCN (Kipf and Welling, 2022), GAT (Veličković et al., 2018), GIN (Xu et al., 2018), Graphormer (Ying et al., 2021) and Graphormer-p for comparisons.

Table 2: Experimental Results on Test-Mini Dataset.

Model	Shot	Method	BACE (152)	HIV (4000)	MUV (4000)	Tox21 (4000)	ToxCast (4000)	BBBP (204)	CYP450 (4000)
Llama3-8b	4-shot	Struct&Sim-CoT	0.6688	0.6393	0.6154	0.6046	0.6440	0.5064	0.6293
		Sim-CoT	0.7225	0.6449	0.5541	0.6600	0.6654	0.5411	0.7176
Qwen2.5-7b			0.7560	**0.6476**	**0.6237**	**0.6642**	**0.7087**	0.5608	0.7279
GPT-4o			**0.7673**	0.6209	0.5486	0.6580	0.6882	**0.6291**	**0.7471**

Metric. We employ ROC-AUC(Area Under the Receiver Operating Characteristic curve) as the evaluation metric for classification tasks, and RMSE (Root Mean Squared Error) for regression tasks evaluation.

4.2 The Outperformance of MolRAG(RQ1)

4.2.1 Performance on Classification Tasks

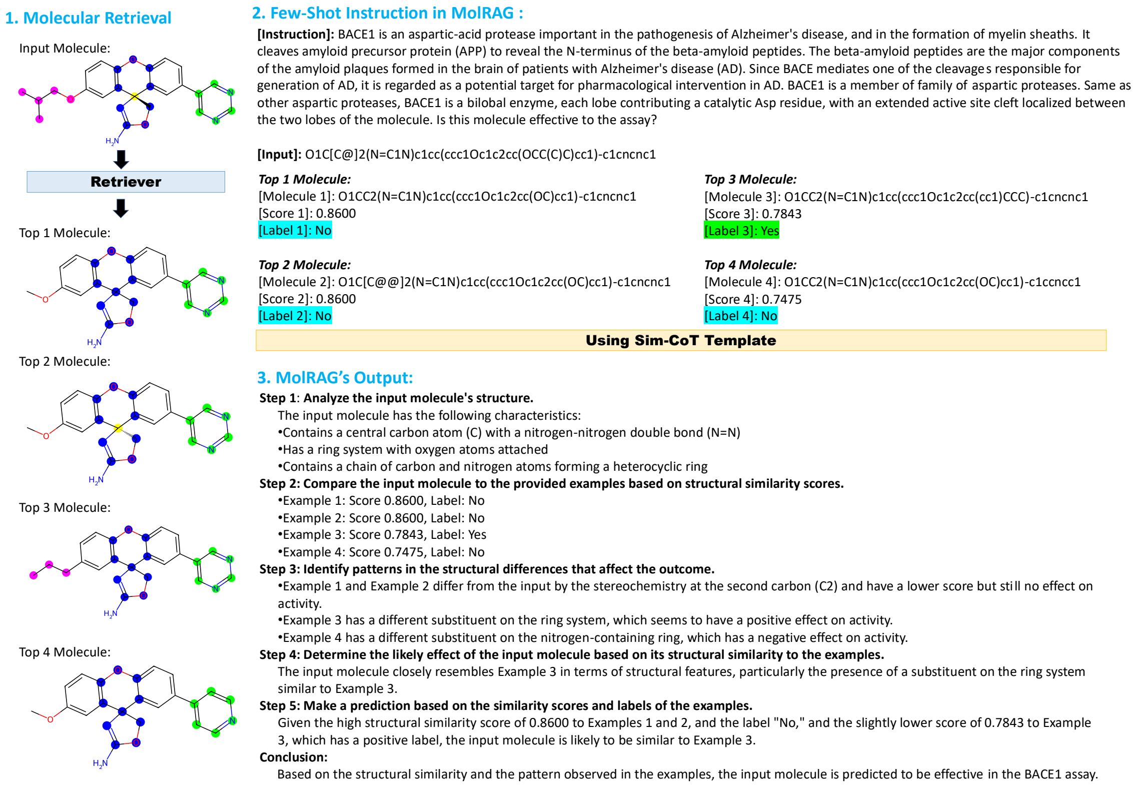
To address this question, we design two experimental settings Direct Answering Setting and Struct-CoT Setting for the General LLM. We choose Llama3-8b-Instruct (Touvron et al., 2023) as the General LLM. In Direct Answering, the model directly answers the question and extracts the final prediction without additional reasoning, while in Struct-CoT setting, the model follows a structured CoT prompting strategy, reasoning step by step before providing an answer. Additionally, we introduce a Random Choice Setting, where the model randomly selects either "Yes" or "No" as the answer. Table 1 shows the performance of these three settings. The result values of Random Choice are very close to the General Model experimental results, indicates that the performance of both the Direct Answering and Strcut-CoT settings, without retrieval, is proximate to random choice. This is because the model lacks domain knowledge and this suggests that current LLMs struggle to perform molecular property prediction without retrieving relevant knowledge. In summary, MolRAG demonstrates performance improvements across all datasets comparing with the General LLM, ranging from 1.1% to 45.7%, highlighting its efficacy and versatility.

To compare with the general LLM performance, we conduct experiments within the MolRAG framework. We evaluate three different retrieval settings: 1-shot, 2-shot, and 4-shot retrieval on Llama3-8b-Instruct (Touvron et al., 2023). The results for MolRAG as shown in Table 1 indicates that regardless of the number of shots retrieved, the performance consistently outperforms the generalist LLM. Specially, under the 4-shot setting of MolRAG, our approach achieves superior results compared to pretrained methods across multiple benchmark datasets, including BACE, CYP450, Tox21, and ToxCast. In particular, on the BACE and ToxCast datasets, our method approaches the performance of supervised methods, highlighting the effectiveness of MolRAG in molecular property prediction tasks. Additionally, Table 1 provides experimental data from both pretrained and supervised methods for reference. A comparison reveals that MolRAG outperforms the pretrained methods across four datasets, and matches the performance of supervised methods on two of these datasets. This demonstrates the versatile effectiveness of MolRAG.

4.2.2 Performance on Regression Tasks

To address this question, we choose Llama3-8b-Instruct as the backbone model and use the Sim-CoT template to conduct experiments. We compare MolRAG with pretrained LLMs and GNN-based methods, predicting solubility (ESOL), free energy of hydration (FreeSolv), and lipophilicity (Lipo). Specifically, considering the higher difficulty of regression tasks, we introduce two additional experimental settings: (1) We include descriptions of the units in the instructions, and (2) We adopt the idea of CoT-SC (Wang et al., 2023), generating multiple answers and selecting the median of these results as our final answer.

Experimental results are listed in Table 3. Results show that MolRAG outperforms Gimlet on Lipo through the proposed framework, but performance on ESOL and FreeSolv remains an area for improvement. We analyze the results and find out the challenges for LLMs to complete regression tasks are: (1) Limited numerical reasoning abilities: Current LLMs have difficulty performing complex arithmetic operations, especially when handling multi-step calculations, floating-point precision, and scientific notation; (2) Weak chemical formula and unit conversion handling: LLMs struggle with chemical equations and unit conversions, often misinterpreting expressions and inaccurately transforming units (e.g., kcal/mol to kJ/mol), leading to significant errors in regression tasks.



2. Few-Shot Instruction in MolRAG :

[Instruction]: BACE1 is an aspartic-acid protease important in the pathogenesis of Alzheimer's disease, and in the formation of myelin sheaths. It cleaves amyloid precursor protein (APP) to reveal the N-terminus of the beta-amyloid peptides. The beta-amyloid peptides are the major components of the amyloid plaques formed in the brain of patients with Alzheimer's disease (AD). Since BACE mediates one of the cleavages responsible for generation of AD, it is regarded as a potential target for pharmacological intervention in AD. BACE1 is a member of family of aspartic proteases. Same as other aspartic proteases, BACE1 is a bilobal enzyme, each lobe contributing a catalytic Asp residue, with an extended active site cleft localized between the two lobes of the molecule. Is this molecule effective to the assay?

[Input]: O1C[C@]2(N=C1N)c1cc(ccc1Oc1c2cc(OCC(C)C)cc1)-c1cncnc1

Top 1 Molecule:
[Molecule 1]: O1CC2(N=C1N)c1cc(ccc1Oc1c2cc(OC)cc1)-c1cncnc1
[Score 1]: 0.8600
[Label 1]: No

Top 2 Molecule:
[Molecule 2]: O1C[C@@]2(N=C1N)c1cc(ccc1Oc1c2cc(OC)cc1)-c1cncnc1
[Score 2]: 0.8600
[Label 2]: No

Top 3 Molecule:
[Molecule 3]: O1CC2(N=C1N)c1cc(ccc1Oc1c2cc(cc1)CCC)-c1cncnc1
[Score 3]: 0.7843
[Label 3]: Yes

Top 4 Molecule:
[Molecule 4]: O1CC2(N=C1N)c1cc(ccc1Oc1c2cc(OC)cc1)-c1ccncc1
[Score 4]: 0.7475
[Label 4]: No

Using Sim-CoT Template

3. MolRAG's Output:

Step 1: **Analyze the input molecule's structure.**
The input molecule has the following characteristics:
- Contains a central carbon atom (C) with a nitrogen-nitrogen double bond (N=N)
- Has a ring system with oxygen atoms attached
- Contains a chain of carbon and nitrogen atoms forming a heterocyclic ring

Step 2: Compare the input molecule to the provided examples based on structural similarity scores.
- Example 1: Score 0.8600, Label: No
- Example 2: Score 0.8600, Label: No
- Example 3: Score 0.7843, Label: Yes
- Example 4: Score 0.7475, Label: No

Step 3: Identify patterns in the structural differences that affect the outcome.
- Example 1 and Example 2 differ from the input by the stereochemistry at the second carbon (C2) and have a lower score but still no effect on activity.
- Example 3 has a different substituent on the ring system, which seems to have a positive effect on activity.
- Example 4 has a different substituent on the nitrogen-containing ring, which has a negative effect on activity.

Step 4: Determine the likely effect of the input molecule based on its structural similarity to the examples.
The input molecule closely resembles Example 3 in terms of structural features, particularly the presence of a substituent on the ring system similar to Example 3.

Step 5: Make a prediction based on the similarity scores and labels of the examples.
Given the high structural similarity score of 0.8600 to Examples 1 and 2, and the label "No," and the slightly lower score of 0.7843 to Example 3, which has a positive label, the input molecule is likely to be similar to Example 3.

Conclusion:
Based on the structural similarity and the pattern observed in the examples, the input molecule is predicted to be effective in the BACE1 assay.

Figure 3: Case Study of MolRAG. The left side of the figure illustrates the visualization of the input molecule and top k molecules. The right side is divided into two sections: the upper part shows the few-shot instruction, while the lower part displays the MolRAG's output result.

4.3 The Ablation Study of MolRAG(RQ2)

4.3.1 The Effect on Retrieved Numbers

To investigate this problem, we compare the results of three different retrieval numbers:1, 2, 4. The experimental results are summarized in Table 1. Our experimental results indicate that the 2-shot setting outperforms the 1-shot setting, while the 4-shot setting further improves upon the 2-shot results. Notably, it suggests that the more retrieved information retrieved, the more beneficial it is for molecular property prediction tasks. All retrieval-enhanced configurations surpass the performance of models without retrieval.

Table 3: Experimental Results on Test Dataset (Regression Tasks).

Model	Shot	Method	ESOL (113)	Lipo (420)	FreeSolv (65)
Llama3-8b	1-shot	Sim-CoT	4.1424	1.2712	18.1224
	2-shot	Sim-CoT	3.4994	1.1676	6.0827
	4-shot	Sim-CoT	3.2806	**1.1251**	6.1923
Pre-training Methods		Gimlet	**1.132**	1.345	**5.103**
		KVPLM	-	-	-
		MoMu	-	-	-
		Galactica	-	-	-
Graph-based Networks		GCN	1.331	0.760	2.119
		GAT	1.253	0.770	2.493
		GIN	1.243	0.781	2.871

4.4 The Effect on Retrieval Mechanism

To determine whether performance gains stem from the LLM's molecular understanding or sample-based retrieval, we conduct KNN-based majority voting experiments. We apply KNN clustering to group structurally similar molecules. For each query molecule, we retrieve its k-nearest neighbors and assign a property label based on majority voting. Experimental results are listed in Table 4. The results show that while the KNN-based method performs well on certain datasets, its effectiveness is constrained. In contrast, MolRAG with Llama3-8B and Qwen2.5-7B consistently outperforms KNN, highlighting the critical role of LLM's molecular understanding and reasoning ability over simple retrieval-based methods.

Therefore, we can conclude that: (1) KNN-based methods show limited effectiveness, performing well only on datasets with lower molecular feature diversity. (2) LLM's molecular understanding is the primary factor driving performance improvements in molecular property prediction.

4.4.1 The Effect on CoT-based Strategy.

To investigate the impact of different in-context learning on molecular property prediction, we conducted experiments on Llama3-8B-Instruct (Tou-

Table 4: Experimental Results on Retrieval Mechanism.

Model	Shot	Method	BACE (152)	HIV (4113)	MUV (25342)	Tox21 (7069)	ToxCast (137215)	BBBP (204)	CYP450 (5669)
Llama3-8b	4-shot	Sim-CoT	0.7225	0.6436	**0.5616**	0.6393	0.6408	0.5411	0.7229
Qwen2.5-7b			0.7560	**0.6680**	0.5441	**0.6602**	**0.6624**	**0.6042**	0.7229
KNN-Based		-	**0.7628**	0.5786	0.5099	0.6095	0.5894	0.5972	**0.7503**

vron et al., 2023) with Struct-CoT and Sim-CoT strategies respectively. These experiments are all performed using 1-shot, 2-shot, and 4-shot retrieval settings. The results are summarized in Table 1, demonstrating that Sim-CoT consistently outperforms Struct-CoT across all datasets except for BBBP dataset. Our results suggest that guiding the model to leverage the dual mechanism of cross-molecule correlation and property continuity constraints enhances its reasoning capabilities. By focusing on structurally similar molecules and their corresponding properties, the model can make more accurate predictions.

To further compare the influence of Struct-CoT and Sim-CoT, we combine the Struct-CoT and Sim-CoT strategies into a **Struct&Sim-CoT** strategy. In this setting, both Struct-CoT and Sim-CoT strategies are provided to the model and evaluated on the Test-Mini dataset. The experimental results are presented in Table 2. The results show that, except for the MUV dataset, the overall performance of Sim-CoT outperforms Struct&Sim-CoT. This indicates that, on most of the test datasets, generating structural information can impact the decision-making process. As the structure information generated by the model may be inaccurate, which can affect the final judgment.

Error Analysis. Error analysis of Sim-CoT and Struct-CoT are provided in Figure 4 (a) and (b). When using Sim-CoT, the model tends to encounter more errors related to the retrieval process, while with Struct-CoT, the errors are more often related to a lack of domain knowledge. Experimental settings of error analysis are shown in Appendix C.

4.4.2 Different Model Platform

To compare the impact of different generalist LLMs, we evaluated Llama3-8B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct (Yang et al., 2024) and GPT-4o (Achiam et al., 2023) on the Test-Mini dataset. The results, shown in Table 2, indicate that Qwen2.5-7B-Instruct outperforms other LLMs on the HIV, MUV, Tox21, ToxCast, and BBBP datasets, while GPT-4o surpasses other LLMs on the BACE, BBBP, and CYP450 datasets. Overall, Qwen2.5-7B-Instruct outperforms GPT-4o, which in turn achieves better performance than Llama-3-8B-Instruct.

Furthermore, we analyze the reasoning processes of different models and observe that model platforms exhibit varying responses to the Sim-CoT strategy. Llama3-8B-Instruct tends to focus primarily on the retrieved content and similarity scores, comparing the input molecule with similar molecules and using their labels to determine the label for the input molecule. However, the final result is often influenced by the model's reasoning process, which can introduce inconsistencies.

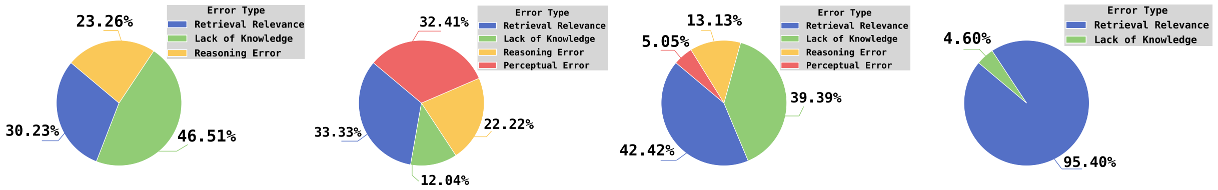
This tendency is even more pronounced in GPT-4o, which almost exclusively relies on the retrieved content to make predictions. When the retrieved content fails to support reasoning toward the correct ground truth, GPT-4o tends to make errors, reflecting the model's strong dependence on the retrieval phase.

In contrast, Qwen2.5-7B-Instruct exhibits a different behavior from the other two models. Even though we only applied the Sim-CoT strategy, Qwen2.5-7B-Instruct still considers the molecular structure and combines it with the similarity scores, enabling more sophisticated reasoning. As a result, Qwen2.5-7B-Instruct's responses are more interpretable and provide more complex insights compared to the other models, offering a deeper understanding of the decision-making process.

Error Analysis. Error analysis of different models are provided in Figure 4 (b), (c) and (d). GPT-4o almost never encounters reasoning errors or perceptual errors. Instead, it tends to have more errors related to the retrieval knowledge which is 95.40%. Qwen2.5-7B-Instruct's errors are primarily concentrated around retrieval and a lack of knowledge, while Llama3-8B-Instruct experiences errors related to both retrieval and perceptual errors in the model. Experimental settings of error analysis are shown in Appendix C.

4.4.3 Different Molecular Retrieval Methods

To evaluate the effect of different molecular retrieval representations, we compared retrieval us-



(a) Struct-CoT on Llama3-8b-Instruct **(b)** Sim-CoT on Llama3-8b-Instruct **(c)** Sim-CoT on Qwen2.5-7b-Instruct **(d)** Sim-CoT on GPT-4o

Figure 4: Error Analysis of different experimental settings of MolRAG. The figures illustrate the types of errors and their corresponding proportions.

ing the Substructure Fingerprint(Daylight Fingerprint) (Stahl and Mauser, 2005) with retrieval using Morgan Fingerprint. The results, summarized in Table 5, indicate that Morgan Fingerprint yield superior performance. However, the use of different fingerprints still produces competitive results, demonstrating that despite variations in molecular representations, MolRAG remains effective. This finding highlights the robustness of the MolRAG framework across different molecular retrieval strategies. Detailed results are shown in Appendix B.1.

4.5 Case Study

Further exploration of one specific test case from the BACE test set provides valuable insights, as shown in Figure 3. This case provides a result based on the 4-shot Sim-CoT template within the MolRAG framework. Given four retrieved molecules, the LLM primarily focuses on the similarity between the input molecule and the retrieved molecules during the reasoning process. In this particular case, three of the retrieved molecules are labeled as “No”, while one is labeled as “Yes”. In most cases, following the provided guidelines, the model predicts the property of the input molecule based on the majority of labels of the retrieved molecules. However, in this case, when the model encounters the top 3 similar molecule, which is labeled as “Yes”, it shifts its decision-making strategy. Instead of relying solely on the majority label, the model begins to analyze and reason about the structural differences between the input molecule and the retrieved examples. This step is reflected in the third part of the reasoning process—“identifies patterns in the structural differences that affect the outcome.” This case study highlights the significance of retrieval-based augmentation and chain-of-thought (CoT) reasoning within the MolRAG framework. By focusing on the structural patterns in molecular property predictions, the model can make more nuanced decisions. More case studies can be found in Appendix D.

5 Conclusion

MolRAG is a retrieval-augmented framework designed for molecular property prediction. To the best of our knowledge, this is the first approach that integrates retrieval-based augmentation with CoT reasoning for this task. Moreover, MolRAG is a training-free method, eliminating the need for additional model fine-tuning. By incorporating relevant information into in-context learning, MolRAG enables LLMs to surpass the performance of pre-trained models and even achieve results comparable to supervised approaches. Additionally, MolRAG enhances interpretability in molecular property prediction, providing researchers with transparent and insightful rationales for model predictions. Our experiments further investigate the impact of different in-context learning strategies, retrieval methods, and generalist LLMs within the MolRAG framework. Experimental results show MolRAG outperforms pre-trained LLMs on four datasets, and even matches supervised methods, achieving performance gains of 1.1%–45.7% over direct prediction approaches, demonstrating versatile effectiveness and robustness.

MolRAG demonstrates strong utility in both industrial and academic settings. It can be directly integrated into workflows, e.g., drug discovery. By supporting local molecular database integration and eliminating the need for task-specific model training, it offers a cost-effective solution that meets the demands of modern pharmaceutical pipelines. MolRAG also shows promise for broader societal and educational impact. With its ability to generate interpretable responses, it supports the dissemination of molecular knowledge to non-experts, and its multi-turn dialogue functionality fosters interactive and personalized engagement with scientific content. As to future work, we intend to integrate embeddings of molecules produced by graph neural networks (Li et al., 2025; Arslan Manzoor et al., 2024; Liu et al., 2024a; Li et al., 2024b) into our model for molecular property prediction.

Limitations

While MolRAG has demonstrated performance improvements in molecular property tasks, it still has some limitations as follows:

- **Static knowledge representation:** The lack of model training limits the continuous evolution of knowledge, as errors in reasoning cannot be used to refine the model's underlying chemical knowledge base.
- **Bounded knowledge scope:** Dependency on predefined molecular databases restricts access to emerging chemical knowledge, lacking dynamic knowledge exploration capabilities akin to literature mining systems.
- **Unverified reasoning pathways:** Logical inconsistencies in structure-property deductions persist without expert-in-the-loop validation, particularly in edge cases requiring nuanced chemical intuition.

Future efforts should focus on developing self-improving frameworks that integrate dynamic knowledge acquisition and expert-guided validation, ultimately advancing LLMs toward chemically rigorous reasoning.

References

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

Muhammad Arslan Manzoor, Ruihong Zeng, Dilshod Azizov, Preslav Nakov, and Shangsong Liang. 2024. Mgm: Global understanding of audience overlap graphs for predicting the factuality and the bias of news media. *arXiv e-prints*, pages arXiv–2412.

Darko Butina. 1999. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *arXiv preprint arXiv:2311.16208*.

Cayque Monteiro Castro Nascimento and André Silva Pimentel. 2023. Do large language models understand chemistry? a conversation with chatgpt. *Journal of Chemical Information and Modeling*, 63(6):1649–1655.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations*.

Jiawei Gu and Shangsong Liang. 2025. Speculative reward model boosts decision making ability of llms cost-effectively. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, et al. 2023. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems*, 36:59662–59688.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Mingyu Jin, Haochen Xue, Zhenting Wang, Boming Kang, Ruosong Ye, Kaixiong Zhou, Mengnan Du, and Yongfeng Zhang. 2024. ProLLM: Protein chain-of-thoughts enhanced LLM for protein-protein interaction prediction. In *First Conference on Language Modeling*.

Mark A Johnson, Gerald M Maggiora, et al. 1990. Concepts and applications of molecular similarity. *(No Title)*.

Thomas N Kipf and Max Welling. 2022. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Guoming Li, Jian Yang, and Shangsong Liang. 2025. Ergnn: Spectral graph neural network with explicitly-optimized rational graph filters. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024a. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6071–6083.

Liang Li, Qisheng Liao, Meiting Lai, Di Liang, and Shangsong Liang. 2024b. Local and global: Text matching via syntax graph calibration. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11571–11575. IEEE.

Ang Liu, Yafang Huang, Yuehong Wu, Siwei Liu, and Shangsong Liang. 2024a. Active learning for multi-class drug-drug interactions prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1611–1615. IEEE.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. 2024b. Conversational drug editing using retrieval and domain feedback. In *The Twelfth International Conference on Learning Representations*.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11.

Siru Ouyang, Zhuosheng Zhang, Bing Yan, Xuan Liu, Yejin Choi, Jiawei Han, and Lianhui Qin. 2024. Structured chemistry reasoning with large language models. In *ICML*.

Martin Stahl and Harald Mauser. 2005. Database clustering with a combination of fingerprint and maximum common substructure methods. *Journal of chemical information and modeling*, 45(3):542–548.

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

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

Andrew D White, Glen M Hocky, Heta A Gandhi, Mehrad Ansari, Sam Cox, Geemi P Wellawatte, Subarna Sasmal, Ziyue Yang, Kangxin Liu, Yuvraj Singh, et al. 2023. Assessment of chemistry knowledge in large language models that generate code. *Digital Discovery*, 2(2):368–376.

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.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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.

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

Haiteng Zhao, Shengchao Liu, Ma Chang, Hannan Xu, Jie Fu, Zhihong Deng, Lingpeng Kong, and Qi Liu. 2023. Gimlet: A unified graph-text model for instruction-based molecule zero-shot learning. *Advances in Neural Information Processing Systems*, 36:5850–5887.

Hongyi Zhou and Jeffrey Skolnick. 2024. Utility of the morgan fingerprint in structure-based virtual ligand screening. *The Journal of Physical Chemistry B*.

A Details of Datasets and Instructions

BACE. The BACE dataset contains molecular structures and their corresponding binding affinities to the Beta-secretase 1 (BACE1) enzyme, which is involved in Alzheimer's disease. It is commonly used to evaluate molecular property prediction models in drug discovery, specifically for compounds that could potentially inhibit BACE1.

HIV. The HIV dataset consists of molecular structures and their activity against the Human Immunodeficiency Virus (HIV). It is used for training machine learning models to predict anti-HIV activity, aiding in the discovery of potential HIV inhibitors for therapeutic applications.

MUV. The MUV dataset is a collection of molecular structures designed for multi-task learning in the context of drug discovery. It includes molecular activity data across multiple targets, which enables the development of models that can predict the activity of a compound against a range of targets simultaneously.

Tox21. The Tox21 dataset contains molecular structures along with data on their toxicity across various biological assays. It is used to develop predictive models for toxicity screening, aiming to identify potentially harmful compounds early in the drug development process.

ToxCast. The ToxCast dataset provides molecular structures and their toxicity data, but it includes a larger set of assays and focuses on environmental chemicals and their potential effects. It is a valuable resource for predicting the toxicological properties of chemicals in a wide range of contexts.

BBBP. The BBBP (Blood-Brain Barrier Permeability) dataset includes molecular structures and their ability to cross the blood-brain barrier. It is used to develop models for predicting the permeability of compounds, which is critical for identifying drug candidates with potential efficacy in treating central nervous system diseases.

CYP450. The CYP450 dataset comprises molecular structures and their inhibitory activities against various human cytochrome P450 enzymes, which are crucial in drug metabolism. This dataset is utilized to develop predictive models for identifying potential drug-drug interactions and assessing metabolic pathways.

ESOL. The ESOL dataset contains compounds with measured water solubility data. The target property is the log solubility in mols per litre, a critical parameter for drug development and formulation. ESOL is known for its relatively small size and low experimental noise, making it a standard benchmark for evaluating molecular regression models.

Lipo. The Lipo dataset contains molecules with recorded experimental values of lipophilicity (logD), which refers to the partition coefficient between octanol and water at a specific pH. Lipophilicity is essential in drug absorption and permeability studies. The dataset is more complex due to its larger size and greater structural diversity, making it suitable for testing model scalability and robustness.

FreeSolv. FreeSolv includes small molecules with experimental and calculated hydration free energies in water. The dataset targets the hydration free energy (kcal/mol), a key property for understanding molecular behavior in aqueous environments. Due to the limited data points and the quantum chemical nature of the property, this dataset presents unique modeling challenges.

Instructions. We adopted the instructions as Gimlet (Zhao et al., 2023). [1]

B Details of Ablation Study

B.1 Ablation on Retrieval Method

Here, we show the setting and results analysis of different molecular retrieval methods.

We use the Substructure Fingerprint as the molecular retrieval representation. The Substructure Fingerprint (Stahl and Mauser, 2005) encodes the presence or absence of specific substructural features, such as functional groups or aromatic rings, within a molecule. It is a binary bit vector, where each bit represents the occurrence of a particular substructure. Substructure fingerprints rely on predefined structural patterns and are limited by the substructure library. In contrast, Morgan fingerprints encode local atomic environments and are more flexible.

The results of Substructure Fingerprint-based experiments are shown in Table 5. Experiment results reveal that similar to the Morgan Fingerprint, the 1-shot approach consistently underperforms compared to the 2-shot method, and the 2-shot approach lags behind the 4-shot method. Notably, while the Substructure Fingerprint-based experiment's performance is not as good as the Morgan Fingerprint, they still yield competitive results.

---

[1] The datasets and instructions are available at https://huggingface.co/datasets/haitengzhao/molecule_property_instruction

Table 5: Experimental Result on the Substructure Fingerprint.

Model	Shot	Method	BACE (152)	HIV (4113)	MUV (25342)	Tox21 (25342)	ToxCast (137215)	BBBP (204)	CYP450 (5669)
Llama3-8b	1-shot	SimCoT	0.6221	0.5759	0.4893	0.5736	0.5630	0.5469	0.6278
	2-shot	SimCoT	0.6449	0.5969	0.5194	0.6185	0.6198	0.5324	0.6862
	4-shot	SimCoT	**0.7164**	**0.6433**	**0.5499**	**0.6375**	**0.6562**	**0.5405**	**0.7045**

C Details of Error Analysis

Error Analysis is provided on the results of four experimental settings: Struct-CoT on Llama3-8b-Instruct, Sim-CoT on Llama3-8b-Instruct, Sim-CoT on Qwen2.5-7b-Instruct, and Sim-CoT on GPT-4o.

For each experimental setting, we sampled a subset of results for error analysis. Errors in each answer have been manually reviewed and classified, and the final statistics are shown in Figure 4.

We categorize reasoning failures into four primary types:

Retrieval Relevance Error: When the retrieved molecules are structurally or functionally dissimilar to the target compound, it can impair the model's ability to make accurate property predictions. This mismatch often disrupts the reasoning process, leading to incorrect conclusions. For example, if the retrieved molecules have low structural similarity to the input molecule, the predictions are more likely to be erroneous.

Lack of Knowledge: The model may fail to recall or apply fundamental chemical principles, such as trends in electronegativity. For example, the LLM may fail to recognize certain functional groups, leading to incorrect reasoning.

Reasoning Error and Perceptual Errpr: These errors come from limitations in the LLM's reasoning and comprehension abilities. For example, we observed that Llama sometimes fail to follow specific instructions, such as: "Here are the examples' structure similarity scores with the input molecule." Instead of utilizing the provided similarity scores, Llama attempts to compute similarity on its own, resulting in hallucinations.

D More Cases

D.1 Case Study on Different Generalist LLMs

Here, we provide the case study of different generalist LLMs. The generalist LLMs are all tested using the Sim-CoT Strategy on the Test-Mini dataset.

Case Study of Llama3-8b-Instruct. As shown in Figure 5, in this case, under the influence of the Sim-CoT strategy, the model first evaluates the similarity of each molecule and makes an initial judgment, as shown by the "similar" and "less similar" evaluations in the case study. Then, it uses the similarity scores to make property predictions for the input molecule.

Case Study of Qwen2.5-7b-Instruct. As shown in Figure 6, in this case, the model influenced by the Sim-CoT strategy, first analyzes the input molecule. It then performs a secondary evaluation based on the structure and scores of the retrieved molecules, ultimately leading to a comprehensive decision by the model.

Case Study of GPT-4o. As shown in Figure 7, in this case, under the influence of the Sim-CoT strategy, the model first analyzes the Molecular Weight, Log P, and Hydrogen Bond Donors and Acceptors from the instruction. It then combines these insights with the similarity scores provided to make a final decision.

D.2 Case Study on Different Strategy

Here, we show the case study of different CoT-based strategies. Different are all tested using the Llama3-8b-Instruct.

Figure 5 shows the case of Sim-CoT. Since this case was mentioned earlier, it will not repeated here.

Case Study of Struct-CoT. As shown in Figure 8, in this case, using the Struct-CoT strategy, Llama3-8b-Instruct first summarizes the structural information of each molecule. It then compares the structure of the input molecule with those of the retrieved molecules and performs an analytical comparison. Finally, it makes a decision regarding the molecular property based on this comparison.

Case Study of Struct&Sim-CoT. As shown in Figure 9, in this case, using the Struct&Sim-CoT strategy, Llama3-8b-Instruct not only summarizes the structural information of each molecule but also evaluates and incorporates the similarity scores.

Figure 5: Case Study: Using Llama3-8b-Instruct and Sim-CoT Template

Case : Using Llama3-8b-Instruct and Sim-CoT Template

Few-shot Instruction:

[Instruction]: Protein kinase A (PKA) is an ubiquitous serine/threonine protein kinase and belongs to the AGC kinase family. It has several functions in the cell, including regulation of immune response, transcription, cell cycle and apoptosis. PKA is a cAMP dependent enzyme that exists in its native inactive form as a 4 subunit enzyme with two regulatory and two catalytic subunits. Binding of cAMP to the regulatory subunit leads to the disassembly of the complex and release of now active catalytic subunits. Is this molecule inhibitor of PKA?

[Input]:COc1ccccc1C1=NOC(Cn2c(=O)c3c(ncn3C)n(C)c2=O)C1

Here are some examples and the structural similarity scores to the input molecule.
Example 1:
[Molecule 1]: COc1ccc(-c2nc3c(c(=O)n(C)c(=O)n3C)n2CC2CC(c3ccccc3)=NO2)cc1OC
[Score 1]: 0.6609
[Label 1]: No
Example 2:
[Molecule 2]: CC(=O)CCCCn1c(=O)c2c(ncn2C)n(C)c1=O
[Score 2]: 0.5591
[Label 2]: No
Example 3:
[Molecule 3]: COc1ccc(C2=NOC(Cn3c(=O)cc(C)c4ccccc43)C2)cc1
[Score 3]: 0.5524
[Label 3]: No
Example 4:
[Molecule 4]: Cn1c(=O)n(CCc2ccccc2)c(=O)c2c1ncn2Cc1ccccc1
[Score 4]: 0.5263
[Label 4]: No

Model Output:

First, let's analyze the input molecule: COc1ccccc1C1=NOC(Cn2c(=O)c3c(ncn3C)n(C)c2=O)C1

Next, I'll compare it to the given examples and their scores:

1. [Molecule 1]: COc1ccc(-c2nc3c(c(=O)n(C)c(=O)n3C)n2CC2CC(c3ccccc3)=NO2)cc1OC
* Score: 0.6609 (similar)
2. [Molecule 2]: CC(=O)CCCCn1c(=O)c2c(ncn2C)n(C)c1=O
* Score: 0.5591 (less similar)
3. [Molecule 3]: COc1ccc(C2=NOC(Cn3c(=O)cc(C)c4ccccc43)C2)cc1
* Score: 0.5524 (less similar)
4. [Molecule 4]: Cn1c(=O)n(CCc2ccccc2)c(=O)c2c1ncn2Cc1ccccc1
* Score: 0.5263 (less similar)

Based on the scores, the input molecule is most similar to Molecule 1 (0.6609), which is labeled as "No" (not an inhibitor of PKA). Since similar structures tend to have similar properties, I'll assume that the input molecule is also likely to be "No" (not an inhibitor of PKA).

Figure 6: Case Study: Using Qwen2.5-7b and Sim-CoT Template.

Case: Using Qwen2.5-7b-Instruct and Sim-CoT Template

Few-shot Instruction:

[Instruction]: Estrogen receptor alpha (ER aplha) is Nuclear hormone receptor. The steroid hormones and their receptors are involved in the regulation of eukaryotic gene expression and affect cellular proliferation and differentiation in target tissues. Ligand-dependent nuclear transactivation involves either direct homodimer binding to a palindromic estrogen response element (ERE) sequence or association with other DNA-binding transcription factors, such as AP-1/c-Jun, c-Fos, ATF-2, Sp1 and Sp3, to mediate ERE-independent signaling. Is this molecule agonists of the estrogen receptor alpha (ER-alpha) signaling pathway?

[Input]:NC(=O)c1ccc[n+]([C@@H]2O[C@H](COP(=O)([O-])OP(=O)(O)OC[C@H]3O[C@@H](n4cnc5c(N)ncnc54)[C@H](O)[C@@H]3O)[C@@H](O)[C@H]2O)c1

Here are some examples and the structural similarity scores to the input molecule.
Example 1:
[Molecule 1]: Nc1ncnc2c1ncn2[C@@H]1O[C@H](COP(=O)(O)OP(=O)(O)OP(=O)(O)O)[C@@H](O)[C@H]1O
[Score 1]: 0.7559
[Label 1]: No

Example 2:
[Molecule 2]: Nc1ncnc2c1ncn2[C@@H]1O[C@H](COP(=O)(O)O)[C@@H](O)[C@H]1O
[Score 2]: 0.7213
[Label 2]: No

Example 3:
[Molecule 3]: Nc1ncnc2c1ncn2[C@@H]1O[C@H](CO)[C@@H](O)[C@H]1O
[Score 3]: 0.6140
[Label 3]: No

Example 4:
[Molecule 4]: CC(C)(COP(=O)([O-])OP(=O)([O-])OC[C@H]1O[C@@H](n2cnc3c(N)ncnc32)[C@H](O)[C@@H]1OP(=O)([O-])O)C(O)C(=O)NCCC(=O)NCCS
[Score 4]: 0.6076
[Label 4]: No

Model Output:

Step 1: Analyze the input molecule's structure.
The input molecule has a complex structure with a nitrogen-containing heterocycle, a carbonyl group, and a phosphate group. It also contains a cyclohexane ring and a phenyl ring.

Step 2: Compare the input molecule with the provided examples.
- Example 1 has a score of 0.7559 and is labeled as "No".
- Example 2 has a score of 0.7213 and is labeled as "No".
- Example 3 has a score of 0.6140 and is labeled as "No".
- Example 4 has a score of 0.6076 and is labeled as "No".

Step 3: Assess structural similarity.
The input molecule shares some common features with the examples, such as the presence of a nitrogen-containing heterocycle and a phosphate group, but it also has unique structural elements that are not present in the examples, such as the specific arrangement of rings and functional groups.

Step 4: Consider the context of estrogen receptor alpha (ER-alpha) agonists.
ER-alpha agonists typically have a specific binding affinity for the receptor and can activate the nuclear transactivation pathway. The examples provided are all labeled as "No", indicating they do not act as ER-alpha agonists.

Step 5: Make a prediction based on structural similarity and known examples.
Given the low structural similarity scores and the consistent labeling of the examples as "No", it is unlikely that the input molecule acts as an ER-alpha agonist.

The model then combines both the structural information and similarity scores to make the final decision regarding the molecular property.

E Details of Prompts

In this section, we provide the details of the main experiment prompts.

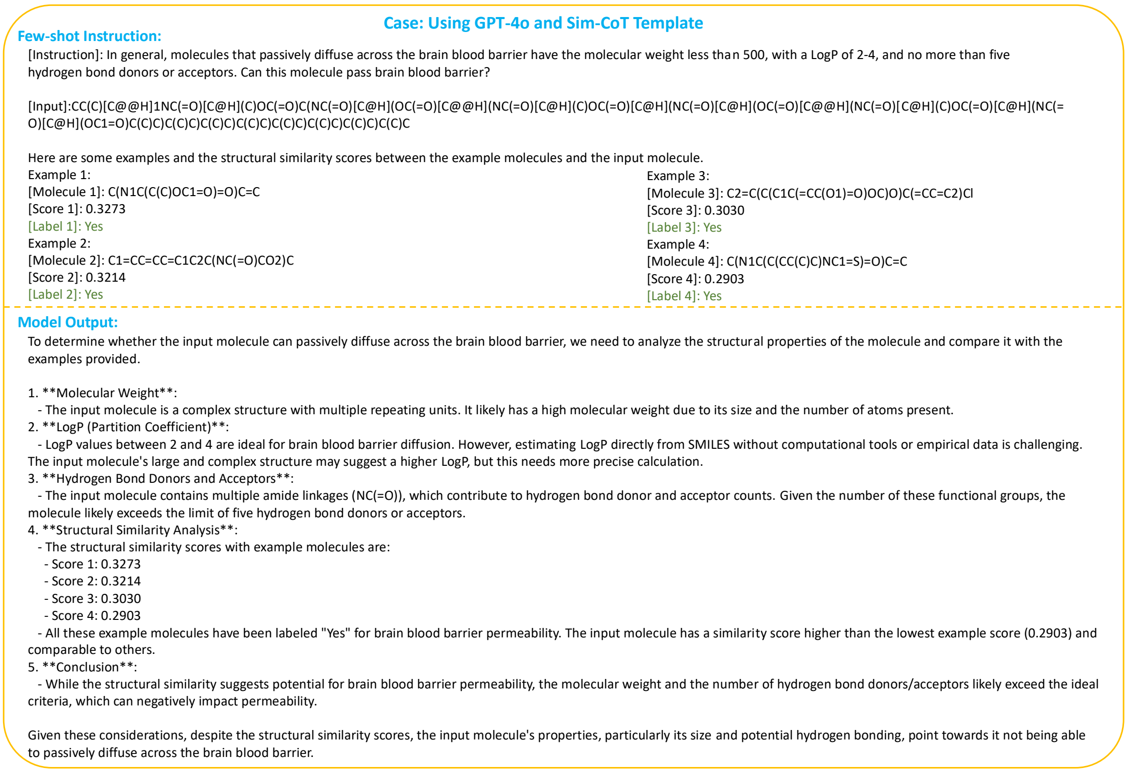
To evaluate the performance of the General LLM and MolRAG, we prompt the LLM to infer the property from the given molecular SMILES string.

Prompts for Direct Answering. We asked the General LLM to answer the question directly, with the prompt described in Table 6.

Prompts for n-shot Struct-CoT. We asked MolRAG to answer the question, with the prompt described in Table 7.

Prompts for n-shot Sim-CoT. We asked MolRAG to answer the question, with the prompt de-

Figure 7: Case Study: Using GPT-4o and Sim-CoT Template.



Case: Using GPT-4o and Sim-CoT Template

Few-shot Instruction:

[Instruction]: In general, molecules that passively diffuse across the brain blood barrier have the molecular weight less than 500, with a LogP of 2-4, and no more than five hydrogen bond donors or acceptors. Can this molecule pass brain blood barrier?

[Input]:CC(C)[C@@H]1NC(=O)[C@H](C)OC(=O)C(NC(=O)[C@H](OC(=O)[C@@H](NC(=O)[C@H](C)OC(=O)[C@H](NC(=O)[C@H](OC(=O)[C@@H](NC(=O)[C@H](C)OC(=O)[C@H](NC(=O)[C@H](OC1=O)C(C)C)C(C)C)C(C)C)C(C)C)C(C)C)C(C)C)C(C)C)C(C)C

Here are some examples and the structural similarity scores between the example molecules and the input molecule.
Example 1:
[Molecule 1]: C(N1C(C(C)OC1=O)=O)C=C
[Score 1]: 0.3273
[Label 1]: Yes
Example 2:
[Molecule 2]: C1=CC=CC=C1C2C(NC(=O)CO2)C
[Score 2]: 0.3214
[Label 2]: Yes
Example 3:
[Molecule 3]: C2=C(C(C1C(=CC(O1)=O)OC)O)C(=CC=C2)Cl
[Score 3]: 0.3030
[Label 3]: Yes
Example 4:
[Molecule 4]: C(N1C(C(CC(C)C)NC1=S)=O)C=C
[Score 4]: 0.2903
[Label 4]: Yes

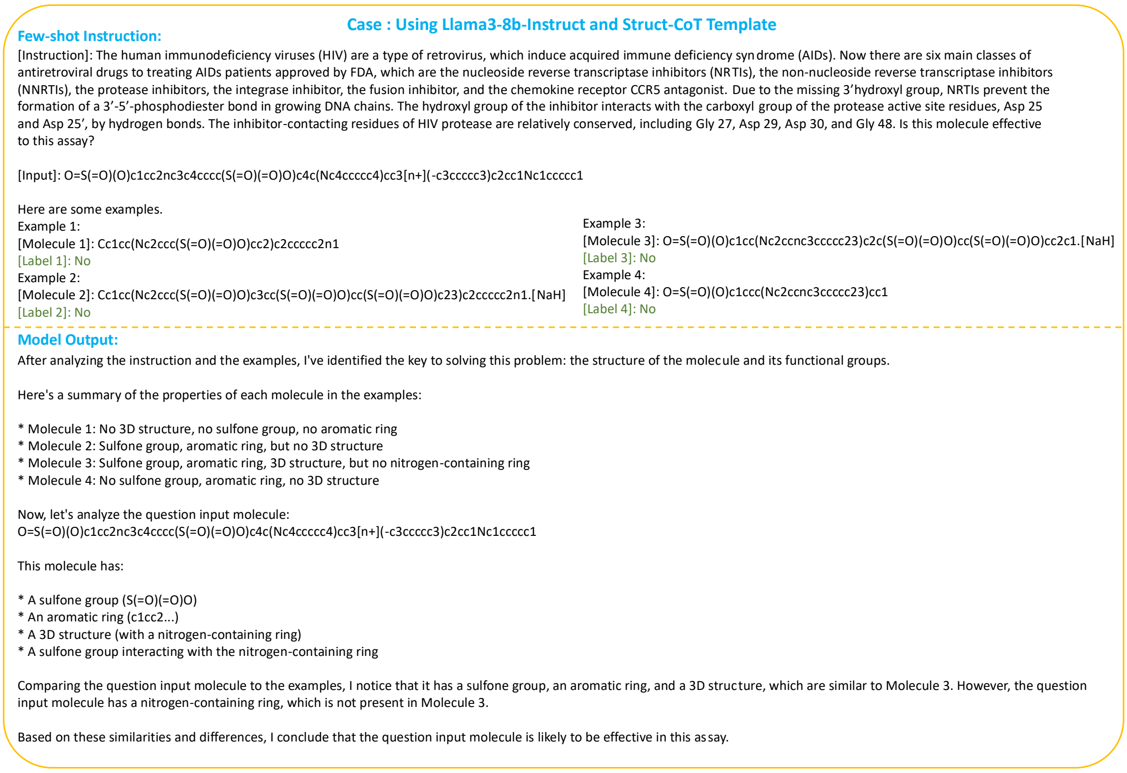
Model Output:

To determine whether the input molecule can passively diffuse across the brain blood barrier, we need to analyze the structural properties of the molecule and compare it with the examples provided.

1. **Molecular Weight**:
  - The input molecule is a complex structure with multiple repeating units. It likely has a high molecular weight due to its size and the number of atoms present.
2. **LogP (Partition Coefficient)**:
  - LogP values between 2 and 4 are ideal for brain blood barrier diffusion. However, estimating LogP directly from SMILES without computational tools or empirical data is challenging. The input molecule's large and complex structure may suggest a higher LogP, but this needs more precise calculation.
3. **Hydrogen Bond Donors and Acceptors**:
  - The input molecule contains multiple amide linkages (NC(=O)), which contribute to hydrogen bond donor and acceptor counts. Given the number of these functional groups, the molecule likely exceeds the limit of five hydrogen bond donors or acceptors.
4. **Structural Similarity Analysis**:
  - The structural similarity scores with example molecules are:
    - Score 1: 0.3273
    - Score 2: 0.3214
    - Score 3: 0.3030
    - Score 4: 0.2903
  - All these example molecules have been labeled "Yes" for brain blood barrier permeability. The input molecule has a similarity score higher than the lowest example score (0.2903) and comparable to others.
5. **Conclusion**:
  - While the structural similarity suggests potential for brain blood barrier permeability, the molecular weight and the number of hydrogen bond donors/acceptors likely exceed the ideal criteria, which can negatively impact permeability.

Given these considerations, despite the structural similarity scores, the input molecule's properties, particularly its size and potential hydrogen bonding, point towards it not being able to passively diffuse across the brain blood barrier.

Figure 8: Case Study: Using Llama3-8b-Instruct and Struct-CoT Template.



Case : Using Llama3-8b-Instruct and Struct-CoT Template

Few-shot Instruction:

[Instruction]: The human immunodeficiency viruses (HIV) are a type of retrovirus, which induce acquired immune deficiency syndrome (AIDs). Now there are six main classes of antiretroviral drugs to treating AIDs patients approved by FDA, which are the nucleoside reverse transcriptase inhibitors (NRTIs), the non-nucleoside reverse transcriptase inhibitors (NNRTIs), the protease inhibitors, the integrase inhibitor, the fusion inhibitor, and the chemokine receptor CCR5 antagonist. Due to the missing 3'hydroxyl group, NRTIs prevent the formation of a 3'-5'-phosphodiester bond in growing DNA chains. The hydroxyl group of the inhibitor interacts with the carboxyl group of the protease active site residues, Asp 25 and Asp 25', by hydrogen bonds. The inhibitor-contacting residues of HIV protease are relatively conserved, including Gly 27, Asp 29, Asp 30, and Gly 48. Is this molecule effective to this assay?

[Input]: O=S(=O)(O)c1cc2nc3c4cccc(S(=O)(=O)O)c4c(Nc4ccccc4)cc3[n+](-c3ccccc3)c2cc1Nc1ccccc1

Here are some examples.
Example 1:
[Molecule 1]: Cc1cc(Nc2ccc(S(=O)(=O)O)cc2)c2ccccc2n1
[Label 1]: No
Example 2:
[Molecule 2]: Cc1cc(Nc2ccc(S(=O)(=O)O)c3cc(S(=O)(=O)O)cc(S(=O)(=O)O)c23)c2ccccc2n1.[NaH]
[Label 2]: No
Example 3:
[Molecule 3]: O=S(=O)(O)c1cc(Nc2ccnc3ccccc23)c2c(S(=O)(=O)O)cc(S(=O)(=O)O)cc2c1.[NaH]
[Label 3]: No
Example 4:
[Molecule 4]: O=S(=O)(O)c1ccc(Nc2ccnc3ccccc23)cc1
[Label 4]: No

Model Output:

After analyzing the instruction and the examples, I've identified the key to solving this problem: the structure of the molecule and its functional groups.

Here's a summary of the properties of each molecule in the examples:

* Molecule 1: No 3D structure, no sulfone group, no aromatic ring
* Molecule 2: Sulfone group, aromatic ring, but no 3D structure
* Molecule 3: Sulfone group, aromatic ring, 3D structure, but no nitrogen-containing ring
* Molecule 4: No sulfone group, aromatic ring, no 3D structure

Now, let's analyze the question input molecule:
O=S(=O)(O)c1cc2nc3c4cccc(S(=O)(=O)O)c4c(Nc4ccccc4)cc3[n+](-c3ccccc3)c2cc1Nc1ccccc1

This molecule has:

* A sulfone group (S(=O)(=O)O)
* An aromatic ring (c1cc2...)
* A 3D structure (with a nitrogen-containing ring)
* A sulfone group interacting with the nitrogen-containing ring

Comparing the question input molecule to the examples, I notice that it has a sulfone group, an aromatic ring, and a 3D structure, which are similar to Molecule 3. However, the question input molecule has a nitrogen-containing ring, which is not present in Molecule 3.

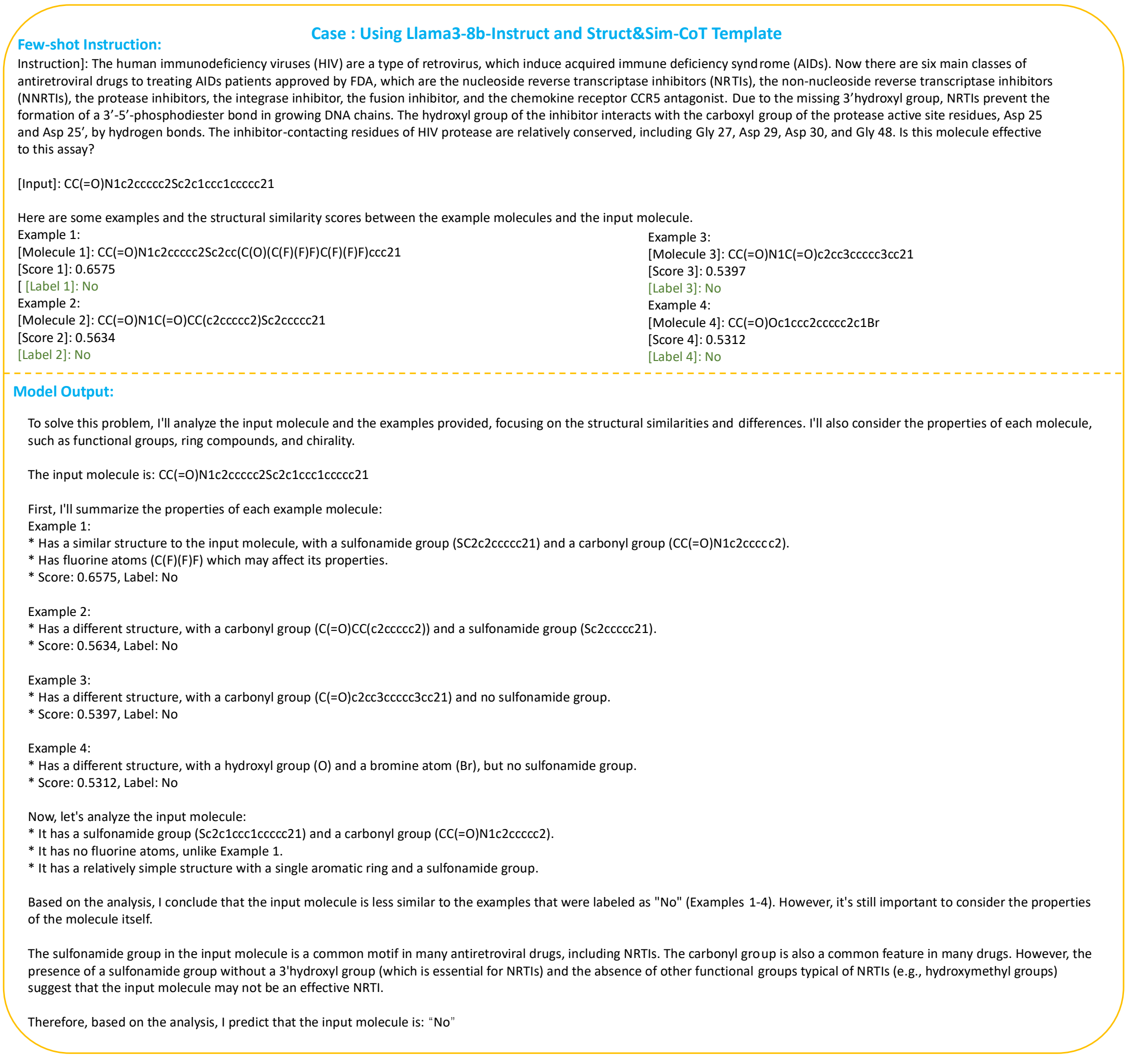
Based on these similarities and differences, I conclude that the question input molecule is likely to be effective in this assay.

scribed in Table 8.

Prompts for n-shot Struct&Sim-CoT. We asked MolRAG to answer the question, with the prompt described in Table 9.

Prompts for regression tasks. We asked MolRAG to answer the question, with the prompt described in Table 10, Table 11 and Table 12.

Figure 9: Case Study: Using Llama3-8b-Instruct and Struct&Sim-CoT Template.



Case : Using Llama3-8b-Instruct and Struct&Sim-CoT Template

Few-shot Instruction:

Instruction]: The human immunodeficiency viruses (HIV) are a type of retrovirus, which induce acquired immune deficiency syndrome (AIDs). Now there are six main classes of antiretroviral drugs to treating AIDs patients approved by FDA, which are the nucleoside reverse transcriptase inhibitors (NRTIs), the non-nucleoside reverse transcriptase inhibitors (NNRTIs), the protease inhibitors, the integrase inhibitor, the fusion inhibitor, and the chemokine receptor CCR5 antagonist. Due to the missing 3’hydroxyl group, NRTIs prevent the formation of a 3’-5’-phosphodiester bond in growing DNA chains. The hydroxyl group of the inhibitor interacts with the carboxyl group of the protease active site residues, Asp 25 and Asp 25’, by hydrogen bonds. The inhibitor-contacting residues of HIV protease are relatively conserved, including Gly 27, Asp 29, Asp 30, and Gly 48. Is this molecule effective to this assay?

[Input]: CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Example 1:
[Molecule 1]: CC(=O)N1c2ccccc2Sc2cc(C(O)(C(F)(F)F)C(F)(F)F)ccc21
[Score 1]: 0.6575
[[Label 1]: No
Example 2:
[Molecule 2]: CC(=O)N1C(=O)CC(c2ccccc2)Sc2ccccc21
[Score 2]: 0.5634
[Label 2]: No

Example 3:
[Molecule 3]: CC(=O)N1C(=O)c2cc3ccccc3cc21
[Score 3]: 0.5397
[Label 3]: No
Example 4:
[Molecule 4]: CC(=O)Oc1ccc2ccccc2c1Br
[Score 4]: 0.5312
[Label 4]: No

Model Output:

To solve this problem, I'll analyze the input molecule and the examples provided, focusing on the structural similarities and differences. I'll also consider the properties of each molecule, such as functional groups, ring compounds, and chirality.

The input molecule is: CC(=O)N1c2ccccc2Sc2c1ccc1ccccc21

First, I'll summarize the properties of each example molecule:
Example 1:
* Has a similar structure to the input molecule, with a sulfonamide group (SC2c2ccccc21) and a carbonyl group (CC(=O)N1c2ccccc2).
* Has fluorine atoms (C(F)(F)F) which may affect its properties.
* Score: 0.6575, Label: No

Example 2:
* Has a different structure, with a carbonyl group (C(=O)CC(c2ccccc2)) and a sulfonamide group (Sc2ccccc21).
* Score: 0.5634, Label: No

Example 3:
* Has a different structure, with a carbonyl group (C(=O)c2cc3ccccc3cc21) and no sulfonamide group.
* Score: 0.5397, Label: No

Example 4:
* Has a different structure, with a hydroxyl group (O) and a bromine atom (Br), but no sulfonamide group.
* Score: 0.5312, Label: No

Now, let's analyze the input molecule:
* It has a sulfonamide group (Sc2c1ccc1ccccc21) and a carbonyl group (CC(=O)N1c2ccccc2).
* It has no fluorine atoms, unlike Example 1.
* It has a relatively simple structure with a single aromatic ring and a sulfonamide group.

Based on the analysis, I conclude that the input molecule is less similar to the examples that were labeled as "No" (Examples 1-4). However, it's still important to consider the properties of the molecule itself.

The sulfonamide group in the input molecule is a common motif in many antiretroviral drugs, including NRTIs. The carbonyl group is also a common feature in many drugs. However, the presence of a sulfonamide group without a 3'hydroxyl group (which is essential for NRTIs) and the absence of other functional groups typical of NRTIs (e.g., hydroxymethyl groups) suggest that the input molecule may not be an effective NRTI.

Therefore, based on the analysis, I predict that the input molecule is: “No”

Table 6: Prompts for Direct-Answering Setting.

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`
Input: `<SMILES>`
Given the input molecule, please answer the question.

Your output must end with a sentence in the format of:
Question Answer:" ".
The value in Question Answer must be EXACTLY one single word `<YES>` or `<NO>`.

Table 7: Prompts for Struct-CoT Setting.

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

(0-SHOT SHOULD NOT INCLUDE) Here are some examples. Examples: `<Examples>`.

Given the input molecule, please answer the question STEP BY STEP. Note:

1. Please use your experience knowledge to analyze the instruction, and find out the key to solve the problem.

2. Please summarize the properties of each molecule in the examples and our question input, The answer to the question may be related to the properties of each molecule, such as: functional group, longest carbon, chain length, aromatic ring, ring compounds, and chirality.
3. Please analyze the similarities and differences between the example molecules and the question input molecule.

Finally, summarize the answer.

Your output must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one single word `<YES>` or `<NO>`.

Table 8: Prompts for Sim-CoT Setting.

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Examples: `<Examples>`.

Based on the above examples and their structure similarity scores with the input molecule, given the input molecule, please answer the question STEP BY STEP. Note:
1. The more similar the molecule, the more similar the properties are, and can be considered as a positive example.

Finally, summarize the answer.

Your output should give an explanation and must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one single word `<YES>` or `<NO>`.

Table 9: Prompts for Struct&Sim-CoT Setting.

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Given the input molecule, please answer the question STEP BY STEP. Note:

1. Please use your experience knowledge to analyze the instruction, and find out the key to solve the problem.

2. The more similar the molecule, the more similar the properties are, and can be considered as a positive example.

3. Please summarize the properties of each molecule in the examples and our question input, the answer to the question may be related to the properties of each molecule, such as: functional group, longest carbon, chain length, aromatic ring, ring compounds, and chirality.
Your output should give an explanation and must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one single word `<YES>` or `<NO>`.

Table 10: Prompts for Regression Tasks(ESOL).

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Examples: `<Examples>`.

Based on the above examples and their structure similarity scores with the input molecule, given the input molecule, please answer the question STEP BY STEP. Note:
1. The more similar the molecules, the more similar the properties are, and can be considered as a positive example.

2. In the calculation formula, S is the intrinsic solubility, P is the octanol/water partition coefficient and MPt is the melting point.
Finally, summarize the answer.

Your output must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one FLOATING-POINT NUMBER.

Table 11: Prompts for Regression Tasks(FreeSolv).

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Examples: `<Examples>`.

Based on the above examples and their structure similarity scores with the input molecule, given the input molecule, please answer the question STEP BY STEP. Note:
1. The more similar the molecules, the more similar the properties are, and can be considered as a positive example.

2. In the calculation formula, $G_{solv,soln}$ represents the solvation free energy of the molecule in solution, $G_{solv,gas}$ is the solvation free energy of the molecule in the gas phase, R is the ideal gas constant, T is the absolute temperature (in Kelvin), pKa is the acid dissociation constant (a measure of the tendency of the molecule to lose a proton).
Finally, summarize the answer.

Your output must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one FLOATING-POINT NUMBER.

Table 12: Prompts for Regression Tasks(Lipo).

Head prompt: You are now working as an excellent expert in chemistry and drug discovery. Given the SMILES representation of a molecule, your job is to predict the property of the molecule. Molecules vary in size, mass, structure, and properties like solubility, toxicity, state, boiling point, hardness, ductility. Their properties depend on structures.

Input prompt: Instruction: `<Instruction>`

Input: `<SMILES>`

Here are some examples and the structural similarity scores between the example molecules and the input molecule.

Examples: `<Examples>`.

Based on the above examples and their structure similarity scores with the input molecule, given the input molecule, please answer the question STEP BY STEP. Note:
1. The more similar the molecules, the more similar the properties are, and can be considered as a positive example.

2. $logD = logP - log(1 + 10^{(pH - pKa)})$. In the calculation formula, P is the octanol/water partition coefficient, pH is a measure of the acidity or basicity of a solution, pKa is the acid dissociation constant (a measure of the tendency of the molecule to lose a proton).
Finally, summarize the answer.

Your output must end with a sentence in the format of:

Question Answer:" ".

The value in Question Answer must be EXACTLY one FLOATING-POINT NUMBER.