MOLE-RAG: Molecular-structure Enhanced Retrieval-Augmented Generation for Chemistry

Anonymous ACL submission

Abstract

Large language models (LLMs) have shown potential for molecular property prediction, but their ability to reason over chemical structures remains limited because molecular representations such as SMILES differ fundamentally from the natural language on which LLMs are primarily trained. To address this semantic and knowledge gap, we propose Molecule-Centric Retrieval-Augmented Generation (MOLE-RAG), a training-free framework for LLM-based molecular property prediction. MOLE-RAG augments each prediction with three complementary sources of inference-time context: retrieved chemistry literature, molecule-specific context (compound synonyms and identifiers, functional group annotations, and physicochemical descriptors), and structurally similar molecules. We evaluate MOLE-RAG across nine molecular property prediction tasks using proprietary, chemistry-specialized, and open-source LLMs. Across general-purpose LLMs, MOLE-RAG improves ROC-AUC by up to 28 points on classification tasks and reduces regression RMSE by up to 67% relative to a SMILES-only baseline. We further find that context source utility varies across models and tasks, with some models benefiting more from textual retrieval, molecular context, or structural retrieval.

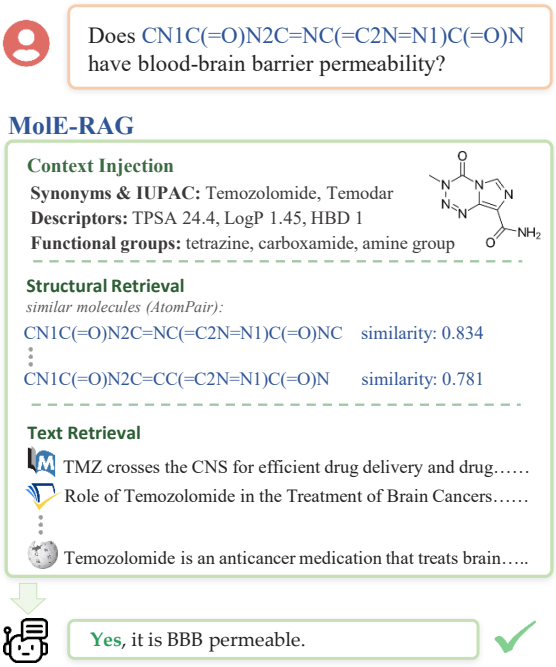

Figure 1: The MOLE-RAG framework illustrated on the BBBP task. Each prediction is augmented with retrieved text passages, structurally similar labeled molecules, and molecule-specific descriptors.

1 Introduction

Molecular property prediction has become increasingly important for accelerating drug discovery by reducing reliance on costly experimental testing (Shen and Nicolaou, 2019). Molecular properties such as toxicity, solubility, permeability, and biological activity are central to drug discovery, as they inform key aspects of candidate quality, including safety, absorption, developability, and therapeutic potential (Segall and Barber, 2014; Lipinski et al., 1997; Schenone et al., 2013). Accurate prediction of these properties can help prioritize promising drug candidates, reduce downstream attrition, and improve the efficiency of molecular screening (Schneider, 2018).

Retrieval-augmented generation (RAG) offers a potential way to address this limitation by providing external evidence at inference time. First, textual retrieval can supply relevant chemical and biomedical knowledge from scientific literature, helping connect molecular inputs to task-specific terminology, known biological mechanisms, and prior experimental evidence. Second, beyond retrieval, molecule-specific context, including structured physicochemical descriptors from cheminformatics toolkits, natural language molecule identifiers, and functional group annotations, can provide information that LLMs may not reliably infer from

SMILES alone (Landrum et al., 2013; Guo et al., 2023b; Tang et al., 2025). Third, structure-based retrieval is motivated by the long-standing use of molecular similarity to relate chemical structure to biological activity, ADME/Tox behavior, and physicochemical properties (Hendrickson, 1991; Bender and Glen, 2004). Recent molecular RAG approaches retrieve structurally similar molecules as contextual examples for LLM-based property prediction (Xian et al., 2025a). Yet it remains unclear whether textual retrieval, molecular context, and structural retrieval consistently improve prediction across diverse molecular property tasks, or how different LLMs behave under these settings.

To address this gap, we propose MOLE-RAG, a Molecule-Centric Retrieval-Augmented Generation framework for LLM-based molecular property prediction. MOLE-RAG integrates three forms of inference-time context: 1) BM25-based textual retrieval (Robertson and Zaragoza, 2009) using LLM-augmented queries with molecule synonyms, IUPAC names, and task-specific vocabulary; 2) molecular context injection, which adds molecule identifiers, functional group annotations, and physicochemical descriptors to the prediction prompt; and 3) structural retrieval using task-adaptive molecular fingerprints to identify structurally similar training molecules. Together, these components bridge the gap between SMILES-based molecular inputs and the chemical evidence needed for property prediction. By providing relevant chemical context at inference time, MOLE-RAG also avoids additional model fine-tuning and post-training.

In this work, we make three main contributions. First, we propose MOLE-RAG, a training-free framework that augments LLM-based molecular property prediction with retrieved chemistry and biomedical literature, molecule-specific context, and structurally similar molecules. Second, we develop a task-adaptive context augmentation strategy that connects SMILES inputs to chemistry literature through LLM-mediated query augmentation, injects interpretable molecule-level identifiers, functional groups, and physicochemical descriptors, and retrieves chemically relevant examples through task-specific structural fingerprints. Third, we systematically evaluate general-purpose, chemistry-specialized, and proprietary LLMs across multiple context configurations to identify when textual retrieval, molecular context, and structural retrieval improve prediction, and how these effects vary across model families and molecular property tasks.

2 Related Works

2.1 LLMs for Chemistry

Since the introduction of transformer-based models, there have been a several works focusing on using language models for science (Cohan et al., 2020; Jin et al., 2023; Zhang et al., 2024b). There has been increasing interest in using LLMs for chemistry, including recent works such as MoleculeNet (Wu et al., 2018a) and MolInstructions (Fang et al., 2023a). ChemDFM (Zhao et al., 2025) is a recent chemistry foundational model that finetunes LLaMa-13B (Touvron et al., 2023). ChemDFM is claimed to outperform general LLMs such as GPT-4 (OpenAI, 2023) and LLaMa-2 (Touvron et al., 2023), but it was only evaluated on ChemLLM-Bench, which is severely limited as its test set is limited to 100 samples per task (Guo et al., 2023a).

2.2 RAG for Property Prediction

Molecular property prediction plays an important role in computational chemistry. By identifying candidate molecules with desired pharmacological properties, accurate molecular property prediction can accelerate drug discovery (Xia et al., 2023; Walters and Barzilay, 2020). Retrieval-Augmented Generation (RAG) has been especially promising because it can leverage the latest advances in general-purpose LLMs without requiring further training. Recent works such as ChemRAG (Zhong et al., 2025a) and MolRAG (Xian et al., 2025b) show the promise of RAG for property prediction, as in-context learning from retrieved examples greatly improve performance.

3 Method

MOLE-RAG is a training-free RAG framework for LLM-based molecular property prediction. Instead of relying on a single source of external context, MOLE-RAG augments each prediction with three complementary forms of inference-time evidence: retrieved chemistry literature, structurally similar molecules, and molecule-specific descriptors. Figure 2 illustrates the overall framework.

3.1 Problem Formulation

Let $D = \{(M_i, y_i)\}_{i=1}^{N}$ be a training set of labeled molecules, where each molecule M_i is represented as a SMILES string with an associated property label y_i. Given a query molecule M, the goal of

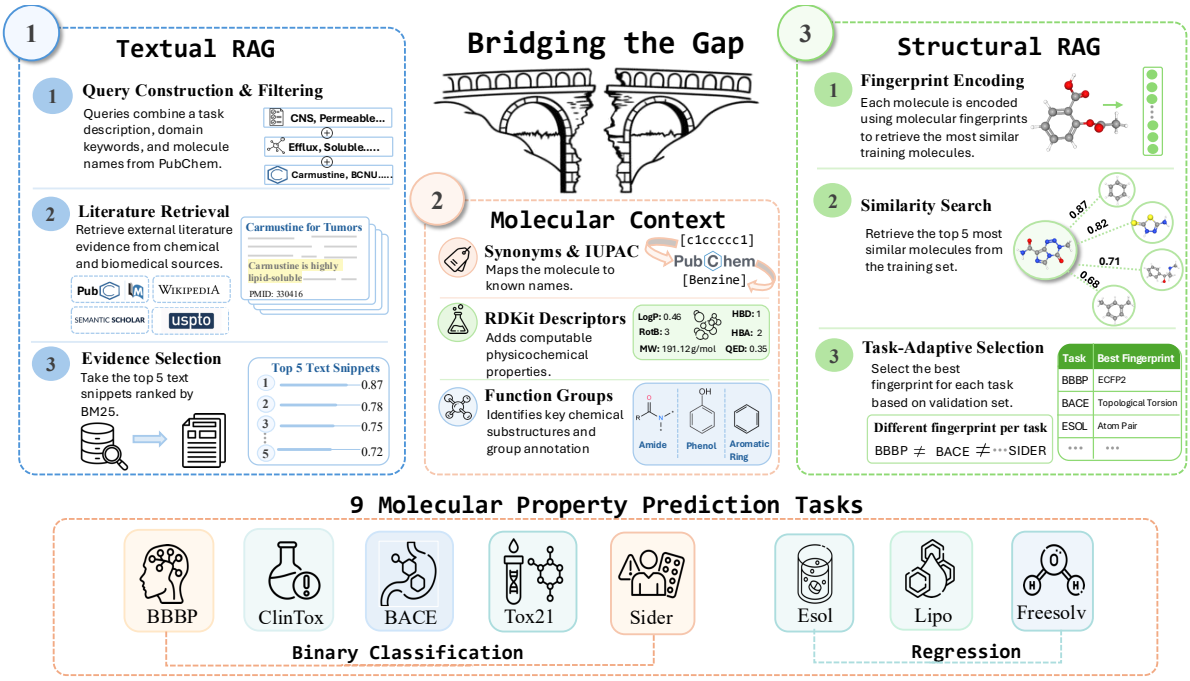

Figure 2: **The MOLE-RAG Framework.** Three complementary sources of inference-time context augment each prediction: (1) **Text retrieval** constructs a hybrid query from the task description, LLM-generated domain keywords, and filtered molecule names, retrieving the top-5 passages from the ChemRAG corpus (Zhong et al., 2025b); (2) **Molecular Context** appends compound identifiers, task-adaptive RDKit descriptors, and functional group annotations; and (3) **Structure retrieval** encodes each molecule using a task-specific molecular fingerprint and retrieves the top-5 most similar training molecules as labeled few-shot examples, where the best fingerprint per task is selected on the validation set. MOLE-RAG is evaluated on nine molecular property prediction tasks.

molecular property prediction is to estimate its task-specific property $\hat{y} = f(M)$, where $\hat{y}$ is either a binary label for classification tasks or a continuous value for regression tasks.

In MOLE-RAG, multiple sources of inference-time context can be used to augment the input molecule. The textual retriever R_{text}, implemented using BM25, retrieves relevant passages from an external chemistry corpus K. The molecular context component derives molecule-specific context directly from M, including molecule synonyms and identifiers, functional group annotations, and RDKit physicochemical descriptors. Finally, the structural retriever R_{struct} performs structure-based retrieval by identifying structurally similar labeled molecules from the training set D. These three context sources are defined as:

$$C_{\text{text}} = R_{\text{text}}(M \mid K),$$
$$C_{\text{mol}} = g_{\text{mol}}(M),$$
$$C_{\text{struct}} = R_{\text{struct}}(M \mid D).$$

Since different LLMs and molecular property tasks may benefit from different types of context, we allow these sources to be used individually or in combination. Let $S \subseteq \{\text{text}, \text{mol}, \text{struct}\}$ denote the selected context configuration for a given experiment. The corresponding augmented context is defined as:

$$C_S = \{C_j : j \in S\},$$

where C_{text}, C_{mol}, and C_{struct} correspond to textual evidence, molecule-specific context, and structurally similar molecules, respectively. The augmented prompt is then constructed as:

$$P_S = (I, M, C_S),$$

where I denotes the task instruction. Finally, the LLM L produces the prediction:

$$\hat{y} = L(P_S).$$

This formulation allows us to evaluate not only whether these inference-time context sources improve molecular property prediction, but also how the usefulness of each context source varies across models, tasks, and prediction settings.

3.2 Textual Retrieval

Given a query molecule M, the textual retriever R_{text} retrieves relevant passages from an external chemistry corpus K. We implement R_{text} using BM25 (Robertson and Zaragoza, 2009) over the ChemRAG corpus (Zhong et al., 2025b), which aggregates text snippets from PubChem, PubMed, USPTO, Semantic Scholar, Wikipedia, and OpenStax textbooks. These sources span molecular databases, biomedical literature, chemical patents, and educational resources. Because BM25 relies on lexical term matching, retrieval quality depends strongly on constructing queries with chemically meaningful terms that overlap with relevant passages.

A key challenge is that SMILES strings are poor lexical queries: they encode molecular structure but share little vocabulary with scientific literature. Therefore, MOLE-RAG uses a task-aware hybrid query that combines the prediction objective, task-specific terminology, and molecule identity:

$$q_{\text{hybrid}} = \big(q_{\text{task}}, q_{\text{keywords}}, q_{\text{mol}}\big).$$

Here, q_{task} denotes the natural language task description, q_{keywords} denotes task-specific chemical terminology, and q_{mol} denotes molecule identity information, such as PubChem synonyms, the IUPAC name, or the SMILES string.

The task description states the prediction objective in natural language, such as *"whether the molecule has blood-brain barrier permeability."* Task keywords are generated once per task using GPT-4o-mini (OpenAI, 2024) and cached for reuse, producing domain-relevant terms such as assay names, pathway terminology, mechanistic keywords, and property-specific synonyms. The molecule identity information is selected from PubChem synonyms when available, with fallback to the IUPAC name or SMILES string.

Together, these components increase lexical overlap with relevant chemistry passages while avoiding noisy identifiers or uninformative tokens. The top-k passages retrieved using q_{hybrid}, with $k = 5$ in our experiments, are prepended to the prediction prompt as textual evidence.

3.3 Molecular Context

Beyond textual and structure-based retrieval, MOLE-RAG directly injects molecule-specific context into the prediction prompt, producing C_{mol} without requiring additional external search. This addresses a limitation of SMILES-only inputs: although SMILES encode molecular topology, they do not explicitly expose chemically meaningful information such as compound identity, functional groups, or physicochemical properties. C_{mol} consists of three components.

Compound identifiers We first inject the molecule's common drug name, up to five additional synonyms, and IUPAC name, sourced from PubChem. These identifiers are selected from the same synonym cache used by R_{text}, with catalog identifiers and registry numbers removed. This grounds the prompt in the molecule's real-world identity, which may help the LLM connect the compound to chemical or pharmacological knowledge encountered during pretraining.

Functional group annotations We detect functional groups using AccFG (Liu et al., 2025) and inject them as a concise natural language statement, such as *"Functional groups present: carboxylic acid, ester, aromatic ring."* Functional groups provide chemically meaningful substructures that are relevant to reactivity, toxicity, and ADME-related properties, but may not be obvious to an LLM from the SMILES string alone.

Physicochemical descriptors We compute RDKit descriptors (Landrum et al., 2013) directly from the SMILES string and inject task-relevant descriptors as a structured block. Rather than including all available RDKit descriptors, which may introduce redundant, weakly relevant, or overly verbose information into the prompt, we select a compact task-specific subset. For each task, descriptors are ranked by their absolute correlation with the training labels, using point-biserial correlation for classification tasks and Pearson correlation for regression tasks, and the top 15 descriptors are retained. This keeps the prompt focused on descriptors most associated with the target property while reducing unnecessary context length.

When applicable, descriptors are accompanied by qualitative labels, such as *"MolLogP: 2.1 (moderate lipophilicity)."* We also append Lipinski's Rule of 5 compliance, which summarizes widely used physicochemical criteria related to oral absorption and permeability, including molecular weight, lipophilicity, hydrogen-bond donors, and hydrogen-bond acceptors (Lipinski et al., 1997). This provides a broadly interpretable drug-likeness signal that is relevant across many small-molecule

Dataset	Classification						Regression		
	BBBP	Tox21	SIDER	ClinTox	BACE	HIV	ESOL	FreeSolv	Lipophilicity
# Molecules	2039	7831	1427	1478	1513	41127	1128	642	4200
# Tasks	1	12	27	2	1	1	1	1	1
MGCN	85.0	70.7	55.2	63.4	73.4	—	1.266	3.349	1.113
GROVER	84.1	80.0	58.2	70.7	83.3	67.5	1.475	3.235	0.974
$\text{MolCLR}_{\text{GCN}}$	73.2	73.5	61.0	85.9	82.6	76.8	1.102	2.241	0.819
$\text{MolCLR}_{\text{GIN}}$	73.5	76.7	60.7	90.4	83.5	77.6	1.091	2.017	0.824
$\text{MolCLR}_{\text{CMPNN}}$	72.4	78.4	59.7	88.0	85.0	77.8	0.911	2.021	0.875
SchNet	84.8	76.6	54.5	71.7	76.6	—	1.045	3.215	0.909
$\text{Llama}_{\text{3.2-3B}}$									
Baseline	48.2	50.0	49.9	48.3	50.4	51.8	3.175	5.995	4.095
+MoLE-RAG	48.3	52.5	50.3	48.8	51.8	58.7	2.181	4.343	1.706
$\text{Mistral}_{\text{7B-Inst}}$									
Baseline	46.1	53.1	50.4	50.0	48.6	45.9	3.474	12.585	3.347
+MoLE-RAG	74.6	54.8	50.8	62.1	76.8	62.4	2.890	4.128	1.210
$\text{Qwen3}_{\text{4B-Inst}}$									
Baseline	53.0	49.6	51.1	50.3	48.9	53.3	2.537	5.524	2.456
+MoLE-RAG	**80.1**	66.2	52.8	56.6	71.1	**73.5**	1.852	2.700	1.105
$\text{ChemDFM}_{\text{14B}}$									
Baseline	79.0	55.0	49.2	49.7	68.2	55.2	4.086	6.523	2.347
+MoLE-RAG	78.1	**68.8**	53.9	59.4	**80.4**	69.2	1.467	3.476	1.021
GPT-4o-mini									
Baseline	54.9	52.3	51.6	52.0	51.3	60.0	2.835	5.475	1.667
+MoLE-RAG	74.7	59.3	53.5	**63.3**	57.4	61.3	1.171	**2.555**	0.964
GPT-5.4-nano									
Baseline	53.4	52.7	51.6	52.8	54.3	50.5	1.723	5.851	1.448
+MoLE-RAG	77.0	67.0	**56.6**	62.7	78.3	65.1	**1.138**	2.880	**0.894**

Table 1: Molecular Property Prediction Benchmark Results. Performance on six classification benchmarks (ROC-AUC) and three regression benchmarks (RMSE). Baselines reported in (Fang et al., 2023b) were included in our evaluation. **Bold dark green** marks the best MoLE-RAG-augmented model per task; underlined light green marks the second-best.

property prediction tasks.

To assess the contribution of each molecular context component, we ablate molecule synonyms and identifiers, functional group annotations, and physicochemical descriptors in Appendix B.

3.4 Structure-Based Retrieval

Given a query molecule M, the structure-based retriever R_{struct} retrieves the top-k most structurally similar molecules from the training set D. We first encode the query molecule and each training molecule $M_i \in D$ using a molecular fingerprint function $\phi(\cdot)$. The similarity between M and M_i is then computed using Tanimoto similarity:

$$s_i = \text{Tanimoto}\big(\phi(M), \phi(M_i)\big).$$

The structure-based context is defined as the top-k labeled molecules with the highest similarity scores:

$$C_{\text{struct}} = \text{TopK}_{(M_i, y_i) \in D}(s_i, k).$$

These retrieved molecules are used as labeled few-shot examples in the prediction prompt. We set $k = 5$ for all experiments. This retrieval strategy is motivated by the principle that structurally similar molecules often exhibit similar biological or physicochemical properties (Bender and Glen, 2004).

We evaluate structure-based retrieval across nine molecular property prediction tasks from MoleculeNet (Wu et al., 2018b), covering six binary classification tasks and three regression tasks. For each task, the training set serves as the retrieval database D, while the validation set is used to select the fingerprint type that yields the best retrieval performance.

Molecular fingerprints encode chemical structures as fixed-length vectors for similarity search (Capecchi et al., 2020). We consider several fingerprint families that capture complementary aspects of molecular structure. Circular fingerprints, including ECFP2, ECFP4, and ECFP6, encode local atom neighborhoods at increasing radii using the Morgan algorithm (Rogers and Hahn, 2010). Functional-class fingerprints, including FCFP4 and FCFP6, represent pharmacophoric atom roles

rather than exact atom identities, allowing similarity to reflect shared functional behavior. MACCS keys encode molecules using a fixed set of predefined structural patterns, while topological fingerprints such as Atom Pair and Topological Torsion capture connectivity patterns and longer-range relationships within the molecular graph (Cereto-Massagué et al., 2015).

Since different fingerprint representations encode distinct structural information and may perform differently across molecular property prediction tasks (Xie et al., 2020), MOLE-RAG does not assume a single fingerprint is optimal across all tasks. Instead, it selects the best-performing fingerprint on the validation set for each task. The selected fingerprint for each task is reported in Appendix A.

3.5 Prompt Design

For each prediction setting, the selected context sources are assembled into a structured prompt. When included, context sources are presented in a fixed order: task instruction, retrieved text passages, molecule-specific context, structurally similar labeled molecules, and the query molecule. This ordering provides broad chemical background first, then molecule-level information, followed by structurally similar examples before the final query.

The output instruction is adapted to the task type, requiring either a categorical prediction for classification tasks or a continuous value for regression tasks.

4 Experiment

4.1 Experiment Settings

Dataset splits We use scaffold splitting with three random seeds and an 8:1:1 training/validation/test ratio across all datasets, following previous studies (Rong et al., 2020; Fang et al., 2023b). The training set is used as the retrieval pool for structure-based retrieval, and the validation set is used for task-adaptive fingerprint selection and molecular descriptor selection.

Models We evaluate MOLE-RAG across LLMs from three categories: proprietary models (GPT-4o-mini(OpenAI, 2024) and GPT-5.4-nano (Singh et al., 2025)); general-purpose open-source models (Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Qwen3-4B-Instruct (Grattafiori et al., 2024; Jiang et al., 2023; Yang et al., 2025)); and the chemistry-specialized model ChemDFM-v2.0-14B (Zhao et al., 2025). All models are evaluated at temperature 0 in a zero-shot setting to ensure reproducibility and isolate the contribution of retrieval augmentation. This diverse model set allows us to examine how model capability and chemistry-specific finetuning affect the benefit of retrieval and molecular context augmentation.

4.2 Baselines

We compare MOLE-RAG against a representative set of supervised and self-supervised graph baselines on the MoleculeNet benchmarks. Supervised graph neural networks include SchNet (Schütt et al., 2017) and MGCN (Lu et al., 2019), which learn directly from molecular graphs under task supervision. Self-supervised pretraining methods include GROVER (Rong et al., 2020) and MolCLR (Wang et al., 2022), the latter evaluated across three GNN backbones (GCN, GIN, CMPNN). Reported baseline numbers are taken from Fang et al. (2023b). Our aim is to characterize the contribution of retrieval and molecular context augmentation to LLM-based prediction rather than to claim state-of-the-art on MoleculeNet, so stronger knowledge-enhanced graph methods such as KANO (Fang et al., 2023b) and GODE (Jiang et al., 2025) are discussed for context but not included as headline comparators.

4.3 Metrics

For classification tasks, we report the mean ROC-AUC across three random seeds. For regression tasks, we report the mean RMSE across three random seeds.

4.4 Effect of MOLE-RAG on Classification Tasks

We evaluate MOLE-RAG on six binary classification tasks: BBBP, BACE, ClinTox, HIV, Tox21, and SIDER. The SMILES-only baseline provides only the molecule SMILES string and task description, while MOLE-RAG augments the prompt with retrieved text, structurally similar labeled molecules, and molecule-specific descriptors. Results are reported as ROC-AUC in Table 1.

Across general-purpose LLMs, MOLE-RAG improves ROC-AUC over the SMILES-only baseline on nearly every dataset, with the largest gains appearing for models whose SMILES-only performance is close to random. Mistral improves on BBBP from 46.1 to 74.6 and on BACE from 48.6 to 76.8 — gains of more than 28 ROC-AUC points.

Qwen3 improves on BBBP from 53.0 to 80.1 and on HIV from 53.3 to 73.5. These results suggest that inference-time molecular context can substantially compensate for the limited chemical information available from SMILES-only prompting.

Without additional context, smaller open-source models generally trail proprietary models on classification. With MOLE-RAG, however, Mistral-7B and Qwen3-4B become competitive with — and on several tasks outperform — the proprietary baselines. Qwen3 + MOLE-RAG obtains the best result among all evaluated LLMs on BBBP (80.1) and HIV (73.5). These findings indicate that access to relevant molecular context is a major bottleneck for smaller LLM-based property prediction, and that retrieval can largely close the gap with stronger proprietary models.

The gains are not limited to smaller open-source models. GPT-5.4-nano + MOLE-RAG is consistently among the top two methods on Tox21, SIDER, ClinTox, and BACE, taking the top score on SIDER (56.6). ChemDFM-14B + MOLE-RAG, the only chemistry-specialized model in our evaluation, achieves the highest ROC-AUC on Tox21 (68.8) and BACE (80.4), and is second-best on BBBP, SIDER, and HIV. The exception is Llama-3.2-3B, which shows only marginal gains (BBBP from 48.2 to 48.3, ClinTox from 48.3 to 48.8) and remains near random on most classification tasks, suggesting that effective use of retrieved chemical context requires sufficient model capacity.

Compared with the supervised and self-supervised graph baselines, MOLE-RAG-augmented LLMs reach competitive but not state-of-the-art performance on classification. On BBBP, Qwen3 + MOLE-RAG (80.1) trails the strongest graph baselines (MGCN at 85.0, SchNet at 84.8) by roughly five ROC-AUC points but exceeds the entire MolCLR family (72.4–73.5). On BACE, ChemDFM + MOLE-RAG (80.4) approaches GROVER (83.3) and the MolCLR family (82.6–85.0). On HIV, Qwen3 + MOLE-RAG (73.5) is within four points of the strongest graph baseline ($\text{MolCLR}_{\text{CMPNN}}$ at 77.8) and clearly exceeds GROVER (67.5). The largest remaining gap is on ClinTox, where the best MOLE-RAG result (GPT-4o-mini at 63.3) lies more than twenty points below $\text{MolCLR}_{\text{GIN}}$ (90.4) — a task that appears to benefit most from explicit graph-level structure.

4.5 Effect of MOLE-RAG on Regression Tasks

We further evaluate MOLE-RAG on three regression tasks: ESOL, FreeSolv, and Lipophilicity. Across all evaluated LLMs, MOLE-RAG consistently lowers regression error relative to the SMILES-only baseline. The largest reductions occur on FreeSolv, where several baselines produce highly inaccurate predictions: Mistral improves from RMSE 12.585 to 4.128 — a 67% reduction — and Qwen3 from 5.524 to 2.700. MOLE-RAG also yields substantial improvements on ESOL (ChemDFM from 4.086 to 1.467, GPT-4o-mini from 2.835 to 1.171) and on Lipophilicity (Llama from 4.095 to 1.706, Mistral from 3.347 to 1.210).

Unlike classification, no single model dominates all regression tasks. GPT-5.4-nano + MOLE-RAG achieves the best result on ESOL (1.138) and Lipophilicity (0.894), while GPT-4o-mini + MOLE-RAG is best on FreeSolv (2.555) and second-best on the other two. This variation suggests that different molecular properties benefit from different combinations of model capability and retrieved context, rather than a single uniformly optimal LLM.

Compared with the graph baselines, MOLE-RAG-augmented LLMs are more competitive on regression than on classification. The best MOLE-RAG results outperform MGCN and GROVER on all three regression tasks, and outperform SchNet on FreeSolv and Lipophilicity. On ESOL, GPT-5.4-nano + MOLE-RAG (1.138) outperforms MGCN (1.266) and GROVER (1.475) but trails SchNet (1.045) and the MolCLR family (0.911–1.102). On FreeSolv, GPT-4o-mini + MOLE-RAG (2.555) outperforms MGCN (3.349), GROVER (3.235), and SchNet (3.215), trailing only the MolCLR family (2.017–2.241). On Lipophilicity, GPT-5.4-nano + MOLE-RAG (0.894) is the lowest RMSE outside the MolCLR family (0.819–0.875), beating SchNet (0.909), GROVER (0.974), and MGCN (1.113). The MolCLR family remains the strongest graph baseline overall but is approached most closely on the regression benchmarks, where targeted descriptor injection appears to be more decisive than message passing.

A second benefit of MOLE-RAG is that it reduces extreme numeric failures from SMILES-only prompting. Several baselines produce regression outputs far outside the expected range, especially on FreeSolv, where Mistral's SMILES-only RMSE of 12.585 is nearly twice that of the next-worst

Method	BBBP	BACE	ClinTox	HIV	Tox21	SIDER
Zero-shot	55.8	53.6	50.5	52.8	52.1	50.6
Naive RAG	49.6	50.4	49.5	52.8	53.4	51.0
MOLE-RAG $_{\text{hybrid}}$	**57.9**	**54.0**	**50.2**	**54.6**	**54.6**	**52.2**

Table 2: Text-retrieval ablation. ROC-AUC averaged across the six evaluated LLMs. Zero-shot and Naive RAG share the HIV value because the SMILES baseline reuses the BM25 pipeline for that task.

Method	BBBP	BACE	ClinTox	HIV	Tox21	SIDER
Zero-shot	55.8	53.6	50.5	52.8	52.1	50.6
Random few-shot	59.7	55.8	54.3	—	51.3	50.3
MOLE-RAG $_{\text{struct}}$	**73.3**	**73.3**	**59.6**	**69.4**	**65.9**	**53.6**

Table 3: Structural-retrieval ablation. ROC-AUC averaged across the six evaluated LLMs. Random few-shot samples 5 training molecules uniformly; MOLE-RAG $_{\text{struct}}$ retrieves the top-5 fingerprint neighbors.

model. Adding retrieved examples and molecule-specific descriptors appears to calibrate the prediction space, reducing both average error and the incidence of extreme outliers. Overall, the regression results show that MOLE-RAG not only improves average accuracy but also makes LLM-based numeric prediction more stable.

4.6 Ablation Variants

To measure the contribution of each retrieval source, we compare MOLE-RAG against a zero-shot baseline and two families of ablation variants. The zero-shot baseline uses only the task instruction and query SMILES. Each ablation variant isolates one form of inference-time context.

Textual retrieval Naive RAG (Lewis et al., 2020; Gao et al., 2023) retrieves the top-5 passages from the ChemRAG corpus using BM25 over the query SMILES. MOLE-RAG $_{\text{hybrid}}$ expands the BM25 query with LLM-generated task keywords and filtered molecule synonyms to test whether task-aware query construction improves retrieval. Table 2 reports the comparison.

Naive BM25 retrieval over the raw SMILES string reduces mean ROC-AUC on BBBP, BACE, and ClinTox and matches the baseline elsewhere, indicating that the SMILES string is a poor lexical query against general chemistry text. The hybrid variant, which augments the BM25 query with LLM-generated task keywords and filtered synonyms, recovers from this regression and yields small but consistent gains over zero-shot on all six tasks. This confirms that query construction matters as much as the corpus itself when retrieving textual context for molecular tasks.

Structural retrieval MOLE-RAG $_{\text{struct}}$ retrieves the top-5 nearest training molecules using the best-performing molecular fingerprint for each task, selected on the validation set, and presents them as labeled in-context demonstrations. We compare this against Random few-shot, which samples 5 training molecules uniformly at random in the same format. Table 3 reports the comparison.

Structural retrieval is the strongest single context source we evaluate. MOLE-RAG $_{\text{struct}}$ improves over the zero-shot baseline on every classification task — by roughly 20 ROC-AUC points on BBBP and BACE, and by 17 points on HIV. Random few-shot serves as a sanity check: it yields modest gains over zero-shot on BBBP, BACE, and ClinTox, but the much larger gap between Random and MOLE-RAG $_{\text{struct}}$ shows that fingerprint-based neighbor selection contributes well beyond the in-context-learning effect of simply seeing labeled molecules.

5 Conclusion

In this work, we introduced MOLE-RAG and benchmarked it against representative supervised and self-supervised graph baselines (MGCN, SchNet, GROVER, and the MolCLR family) on MoleculeNet. We found that retrieval-augmented LLMs close most of the gap to these specialized graph models on regression and on three of six classification tasks, while graph-based methods remain stronger on toxicity benchmarks where graph-level structure carries the most signal. Across model scales, mid-sized open-source LLMs such as Mistral-7B and Qwen3-4B gain the most from MOLE-RAG and become competitive with proprietary baselines, while the smallest model in our evaluation, Llama-3.2-3B, shows only marginal gains, suggesting that effective use of inference-time chemical context requires sufficient model capacity. Together, these findings suggest that, equipped with the right context, training-free LLMs can serve as a practical alternative for molecular property prediction.

6 Limitations

A limitation of this work is that although we used the ChemRAG corpus and BM25 retriever in our ablation study, the significant computational overhead of the large ChemRAG corpus prevented

us from evaluating the full ChemRAG method with dense retrievers. Additionally, we did not evaluate chain-of-thought prompting, which prior work (Xian et al., 2025a) has shown can further improve LLM reasoning on molecular property prediction; integrating CoT with our retrieval pillars is left to future work.

References

Andreas Bender and Robert C Glen. 2004. Molecular similarity: a key technique in molecular informatics. *Organic & biomolecular chemistry*, 2(22):3204–3218.

Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. 2020. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of cheminformatics*, 12(1):43.

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

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023a. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *ArXiv*, abs/2306.08018.

Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. 2023b. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1):32.

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

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhengwen Liang, Zhichun Guo, N. Chawla, O. Wiest, and Xiangliang Zhang. 2023a. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in Neural Information Processing Systems 36*.

Taicheng Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh Chawla, Olaf Wiest, Xiangliang Zhang, and 1 others. 2023b. What can large language models do in chemistry? a comprehensive benchmark on eight tasks. *Advances in neural information processing systems*, 36:59662–59688.

James B Hendrickson. 1991. Concepts and applications of molecular similarity. *Science*, 252(5009):1189–1190.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, John Wilbur, and Zhiyong Lu. 2023. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *ArXiv*.

Greg Landrum and 1 others. 2013. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8(31.10):5281.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25.

Xuan Liu, Sarathkrishna Swaminathan, Dmitry Zubarev, Brandi Ransom, Nathaniel Park, Kristin Schmidt, and Huimin Zhao. 2025. Accfg: Accurate functional group extraction and molecular structure comparison. *Journal of Chemical Information and Modeling*, 65(16):8593–8602.

Chengqiang Lu, Qi Liu, Chao Wang, Zhenya Huang, Peize Lin, and Lixin He. 2019. Molecular property prediction: A multilevel quantum interactions modeling perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1052–1060.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*, volume 4. Now Publishers Inc.

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

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571.

Monica Schenone, Vlado Dančík, Bridget K Wagner, and Paul A Clemons. 2013. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology*, 9(4):232–240.

Gisbert Schneider. 2018. Automating drug discovery. *Nature reviews drug discovery*, 17(2):97–113.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Matthew D Segall and Chris Barber. 2014. Addressing toxicity risk when designing and selecting compounds in early drug discovery. *Drug discovery today*, 19(5):688–693.

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

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

Xiang Tang, Qichang Zhao, Jianxin Wang, and Guihua Duan. 2025. Molfcl: predicting molecular properties through chemistry-guided contrastive and prompt learning. *Bioinformatics*, 41(2):btaf061.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

W Patrick Walters and Regina Barzilay. 2020. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. 2018a. Moleculenet: A benchmark for molecular machine learning. *Preprint*, arXiv:1703.00564.

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

Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z Li. 2023. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. *Advances in Neural Information Processing Systems*, 36:64774–64792.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025a. 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.

Ziting Xian, Jiawei Gu, Lingbo Li, and Shangsong Liang. 2025b. 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, Vienna, Austria. Association for Computational Linguistics.

Liangxu Xie, Lei Xu, Ren Kong, Shan Chang, and Xiaojun Xu. 2020. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Frontiers in pharmacology*, 11:606668.

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*.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, and 1 others. 2024a. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024b. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Conference on Empirical Methods in Natural Language Processing*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu,

Su Zhu, and 1 others. 2025. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4).

Xianrui Zhong, Bowen Jin, Siru Ouyang, Yanzhen Shen, Qiao Jin, Yin Fang, Zhiyong Lu, and Jiawei Han. 2025a. Benchmarking retrieval-augmented generation for chemistry. *ArXiv*, abs/2505.07671.

Xianrui Zhong, Bowen Jin, Siru Ouyang, Yanzhen Shen, Qiao Jin, Yin Fang, Zhiyong Lu, and Jiawei Han. 2025b. Benchmarking retrieval-augmented generation for chemistry. *arXiv preprint arXiv:2505.07671*.

Appendix organization The appendix is organized as follows. Appendix A reports the task-specific molecular fingerprints selected for structure-based retrieval. Appendix B presents the RDKit descriptors selected for molecule-specific context injection. Appendix C provides additional results for multitask datasets. Appendix D describes the use of AI assistants, and Appendix E summarizes the scientific artifacts used in this work. Finally, Appendix F presents the prompt templates used across experimental conditions.

A Best Fingerprint Per Task

To select the most effective structure-based retrieval representation, we performed a validation sweep over multiple molecular fingerprint types for each task. The training set was used as the retrieval pool, and each validation molecule was matched to its top-k most similar training molecules under each fingerprint representation. For classification tasks, we selected the fingerprint with the highest validation ROC-AUC, using accuracy as a secondary criterion when applicable. For regression tasks, we selected the fingerprint with the lowest validation RMSE, using MAE as a secondary criterion when applicable. The selected fingerprint was then fixed and used for test-set evaluation.

The selected fingerprints varied across tasks, suggesting that no single molecular representation was uniformly optimal across all property prediction settings. Circular fingerprints such as ECFP2 were selected for BBBP, indicating that local atom-neighborhood patterns were useful for blood-brain barrier permeability prediction. In contrast, topological representations such as Topological Torsion and Atom Pair were selected for BACE, ESOL, Lipo, and Tox21, suggesting that longer-range connectivity patterns can be informative for activity, solubility, lipophilicity, and toxicity-related tasks. Functional-class fingerprints such as FCFP2 were selected for ClinTox and SIDER, indicating that pharmacophoric or functional-role information may be useful for clinical toxicity and side-effect prediction. Overall, these results support the use of task-adaptive fingerprint selection rather than relying on a single fixed structural representation for all datasets.

Table 4 reports the selected fingerprint for each task.

B Best RDKit Descriptors Per Task

For each task, we retained the top 15 RDKit descriptors ranked by absolute correlation with the target label. For classification tasks, descriptors were ranked using point-biserial correlation, while regression tasks used Pearson correlation. For multi-assay datasets, descriptor ranking was performed separately for each assay.

Across tasks, we observed several recurring descriptor patterns. Molecular size and complexity descriptors, such as `MolWt`, `HeavyAtomCount`, `BertzCT`, and Chi descriptors, appeared frequently across both classification and regression tasks, suggesting that global molecular structure provides useful signal for many property prediction settings. Polarity and hydrogen-bonding descriptors, including `TPSA`, `NOCount`, `NumHDonors`, `NumHAcceptors`, and `NHOHCount`, were especially prominent for BBBP and FreeSolv, which is consistent with the importance of polar surface area and hydrogen-bonding capacity for permeability and solvation-related properties. Lipophilicity and surface-area descriptors, such as `MolLogP`, `SlogP_VSA`, and `SMR_VSA`, also appeared across multiple tasks, reflecting the role of hydrophobicity and surface-area partitioning in molecular activity, toxicity, solubility, and lipophilicity. Finally, several functional-group descriptors, including `fr_azo`, `fr_quatN`, `fr_aniline`, `fr_phenol`, and `fr_COO`, were highly ranked for specific datasets, indicating that task-specific substructures provide complementary signal beyond general physicochemical descriptors.

Tables 5 and 6 report the top 15 RDKit descriptors for classification tasks. Table 7 reports the top 15 RDKit descriptors for regression tasks.

C Results for Multitask Datasets

Tables 8, 9, and 10 report results for the multitask binary classification datasets. We report ROC-AUC for the SMILES-only baseline and the full MOLE-RAG setting. These results summarize performance over the available task labels within each dataset.

Across multitask datasets, MOLE-RAG generally improves ROC-AUC over the SMILES-only baseline for most models. The gains are especially clear on ClinTox, where nearly all evaluated models improve with MOLE-RAG; for example, Mistral increases from 0.500 to 0.621, ChemDFM from 0.497 to 0.594, GPT-4o-mini from 0.520 to 0.633, and GPT-5.4-nano from 0.528 to 0.627. On

Dataset	Selected Fingerprint
BACE	Topological Torsion
BBBP	ECFP2
ClinTox	FCFP2
ESOL	Atom Pair
FreeSolv	MACCS
HIV	RDKit
Lipo	Atom Pair
SIDER	FCFP2
Tox21	Atom Pair

Table 4: Selected structure-based retrieval fingerprints for each molecular property prediction task. Fingerprints were selected using validation-set performance and then fixed for test-set evaluation.

Rank	BACE	BBBP	ClinTox
1	NumHeteroatoms	TPSA	BalabanJ
2	BertzCT	NOCount	NumAromaticHeterocycles
3	HeavyAtomCount	NumHDonors	fr_quatN
4	Chi0	NumHeteroatoms	fr_NH0
5	Chi1	NHOHCount	fr_Ar_N
6	SlogP_VSA2	NumHAcceptors	fr_aniline
7	MaxEStateIndex	VSA_EState3	NumAromaticRings
8	MaxAbsEStateIndex	PEOE_VSA10	MaxPartialCharge
9	HeavyAtomMolWt	EState_VSA10	SMR_VSA3
10	NumValenceElectrons	PEOE_VSA1	BCUT2D_LOGPLOW
11	ExactMolWt	qed	BCUT2D_MRLOW
12	MolWt	EState_VSA1	PEOE_VSA8
13	LabuteASA	SMR_VSA1	SlogP_VSA8
14	SMR_VSA1	fr_lactam	SlogP_VSA10
15	Chi0n	SlogP_VSA2	PEOE_VSA3

Table 5: Top 15 RDKit descriptors for BACE, BBBP, and ClinTox.

Rank	HIV	Tox21	SIDER
1	fr_azo	NumAromaticCarbocycles	AvgIpc
2	NumHeteroatoms	fr_benzene	FpDensityMorgan3
3	TPSA	NumAromaticRings	MaxAbsEStateIndex
4	BertzCT	SMR_VSA7	MaxEStateIndex
5	HeavyAtomMolWt	SlogP_VSA6	BCUT2D_CHGHI
6	ExactMolWt	SlogP_VSA8	BCUT2D_LOGPHI
7	MolWt	Chi3v	NumAromaticRings
8	NOCount	RingCount	fr_ketone
9	SMR_VSA10	BertzCT	NumAliphaticCarbocycles
10	PEOE_VSA13	Chi4v	fr_NH0
11	LabuteASA	fr_phenol_noOrthoHbond	RingCount
12	Kappa1	VSA_EState6	NumSaturatedCarbocycles
13	Chi0	fr_phenol	qed
14	HeavyAtomCount	NumAliphaticCarbocycles	fr_piperzine
15	NumValenceElectrons	NumSaturatedCarbocycles	PEOE_VSA3

Table 6: Top 15 RDKit descriptors for HIV, Tox21, and SIDER.

Tox21, MOLE-RAG also improves most evaluated models, with particularly large gains for Qwen, ChemDFM, and GPT-5.4-nano. SIDER shows more modest gains overall, although GPT-5.4-nano, ChemDFM, GPT-4o-mini, and Qwen still improve relative to the SMILES-only baseline. These results suggest that multitask datasets benefit from inference-time molecular context, but the magnitude of improvement depends on both the model and the specific set of assay labels.

D Usage of AI Assistants

In preparing this work, we used AI-based writing assistants to improve sentence structure, correct grammatical errors, and enhance overall readability. These tools were used only for language editing and presentation refinement. They did not determine the research questions, experimental design, methodology, implementation, analysis, or conclusions of the paper. All scientific claims, experimental results, and interpretations were reviewed and verified by the authors.

We also used large language models (LLMs) as part of the experimental pipeline. Specifically, GPT-4o-mini was used to generate task-specific keywords for the hybrid textual retrieval queries. These generated keywords were cached and reused across experiments to improve lexical overlap with chemistry-related passages. In addition, several LLMs, including proprietary, open-source, and chemistry-specialized models, were evaluated as prediction models in our benchmark. Their use as experimental subjects is described in the main method and experiment sections.

No AI assistant was used to fabricate data, alter experimental results, or make autonomous scientific decisions. All generated text, prompts, retrieval outputs, and model predictions were inspected, processed, and analyzed by the authors according to the experimental protocol.

E Scientific Artifacts

This work uses publicly available datasets, software libraries, pretrained models, and retrieval corpora. All artifacts were used in a manner consistent with their intended research purposes.

Datasets We evaluate on molecular property prediction tasks from MoleculeNet (Wu et al., 2018a), including BBBP, BACE, ClinTox, HIV, Tox21, SIDER, ESOL, FreeSolv, and Lipophilicity. These datasets are standard benchmarks for molecular

Rank	ESOL	FreeSolv	Lipo
1	MolLogP	TPSA	MolLogP
2	PEOE_VSA6	NumHDonors	fr_COO
3	MolMR	NOCount	fr_COO2
4	Chi0v	NHOHCount	NumAromaticRings
5	LabuteASA	MinPartialCharge	RingCount
6	MolWt	PEOE_VSA1	fr_Al_COO
7	HeavyAtomMolWt	NumHAcceptors	BalabanJ
8	ExactMolWt	MaxAbsPartialCharge	FpDensityMorgan1
9	Chi1v	SlogP_VSA2	AvgIpc
10	BCUT2D_LOGPHI	MaxAbsEStateIndex	SlogP_VSA6
11	Chi2v	MaxEStateIndex	VSA_EState6
12	FpDensityMorgan1	VSA_EState2	SMR_VSA7
13	Chi1	VSA_EState3	NHOHCount
14	HeavyAtomCount	SMR_VSA3	BertzCT
15	Chi3v	MinAbsPartialCharge	EState_VSA7

Table 7: Top 15 RDKit descriptors for regression tasks. Descriptors were ranked by absolute Pearson correlation with the target value.

Model	SMILES ROC-AUC	MOLE-RAG ROC-AUC
Llama-3.2-3B-Instruct	0.483	0.488
Mistral-7B-Instruct-v0.3	0.500	0.621
Qwen3-4B-Instruct-2507	0.503	0.566
ChemDFM-v2.0-14B	0.497	0.594
GPT-4o-mini	0.520	0.633
GPT-5.4-nano	0.528	0.627

Table 8: ROC-AUC results on ClinTox.

Model	SMILES ROC-AUC	MOLE-RAG ROC-AUC
Llama-3.2-3B-Instruct	0.500	0.525
Mistral-7B-Instruct-v0.3	0.531	0.548
Qwen3-4B-Instruct-2507	0.496	0.662
ChemDFM-v2.0-14B	0.550	0.688
GPT-4o-mini	0.523	0.593
GPT-5.4-nano	0.527	0.670

Table 9: ROC-AUC results on Tox21.

Model	SMILES ROC-AUC	MOLE-RAG ROC-AUC
Llama-3.2-3B-Instruct	–	0.503
Mistral-7B-Instruct-v0.3	0.504	0.508
Qwen3-4B-Instruct-2507	0.511	0.528
ChemDFM-v2.0-14B	0.492	0.539
GPT-4o-mini	0.516	0.535
GPT-5.4-nano	0.516	0.566

Table 10: ROC-AUC results on SIDER.

property prediction and were used only for evaluating model performance under scaffold-based train/validation/test splits.

Retrieval corpus For textual retrieval, we use the ChemRAG corpus (Zhong et al., 2025a), which aggregates chemistry-related text from resources such as PubChem, PubMed, USPTO, Semantic Scholar, Wikipedia, and OpenStax textbooks. We use this corpus only as an external retrieval source for inference-time textual evidence.

Software tools We use RDKit to compute molecular descriptors, molecular fingerprints, Tanimoto similarity, and other cheminformatics features. RDKit (Landrum et al., 2013) was used for molecular structure analysis and descriptor computation, consistent with its intended use. We also use AccFG (Liu et al., 2025) for functional group annotation, which provides molecule-level functional group information included in the molecular context prompt.

Models We evaluate proprietary, open-source, and chemistry-specialized LLMs, including GPT-4o-mini (OpenAI, 2024), GPT-5.4-nano (Singh et al., 2025), Llama-3.2-3B-Instruct (Touvron et al., 2023), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), Qwen3-4B-Instruct (Yang et al., 2025), ChemLLM-7B-Chat (Zhang et al., 2024a), and ChemDFM-v2.0-14B (Zhao et al., 2025). These models were used as inference-time predictors and were not fine-tuned in this work. All models were evaluated under the same prompt-based framework with temperature set to 0.

Package details All experiments were implemented in Python 3.10.12. LLM inference for local models used PyTorch 2.8.0 with CUDA 12.8, Hugging Face Transformers 4.57.1, Accelerate 1.13.0, Datasets 2.19.2, and Tokenizers 0.22.1. Molecular preprocessing, descriptor computation, and fingerprint-based similarity search used RDKit 2026.03.2. Functional group annotations were generated using AccFG. BM25 retrieval was implemented using `rank_bm25`. Data processing used NumPy 2.2.6 and pandas 2.3.3. API-based model calls used the OpenAI Python package 2.38.0. We also used SentencePiece 0.2.1 and Protobuf 7.35.0 for tokenizer and model-loading support.

License and intended use All datasets, corpora, software tools, and pretrained models were used according to their respective licenses and terms of

use. The artifacts were used for molecular property prediction, retrieval, descriptor computation, and benchmarking, which are consistent with their intended research applications. No dataset was used to identify individuals or make clinical decisions.

F Prompt Templates

This section provides representative prompt templates used in our experiments. To fit the two-column format, we present compact templates that preserve the main instruction structure and output constraints. Context blocks were included or omitted depending on the experimental condition. The SMILES-only baseline used only the task instruction and query molecule. The single-component ablation prompts isolate textual retrieval, molecular context, and structure-based retrieval, respectively. The full MOLE-RAG prompt combines retrieved chemistry passages, molecule-specific context, and structurally similar labeled molecules. We also include the synonym-filtering prompt used to construct the filtered synonym cache.

Prompt 1: SMILES-Only Classification

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

SMILES: [QUERY_SMILES]

Reply with EXACTLY ONE WORD:
either 'Yes' or 'No'.
Output nothing else. No explanation,
no punctuation, no context.
```

Prompt 2: SMILES-Only Regression

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

SMILES: [QUERY_SMILES]

Predict the [REGRESSION_PROPERTY].
Valid predictions are real numbers
[VALID_RANGE].

Output EXACTLY ONE number
(example: [EXAMPLE_VALUE]).
No units, no explanation.

ANSWER:
```

Prompt 3: Textual Retrieval Only

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

Retrieved chemistry context:
[TOP-5_RETRIEVED_PASSAGES]

Molecule:
SMILES: [QUERY_SMILES]

Reply with EXACTLY ONE WORD:
either 'Yes' or 'No'.
Output nothing else. No explanation,
no punctuation, no context.
```

Prompt 4: Molecular Context Only

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

Compound identifiers:
Name: [COMMON_NAME]
Other names: [SYNONYMS]
IUPAC: [IUPAC_NAME]

Functional groups:
[FUNCTIONAL_GROUPS]

RDKit descriptors:
[DESCRIPTOR_1]: [VALUE_1]
...
[DESCRIPTOR_15]: [VALUE_15]

Lipinski Rule of 5:
MolWt=[VALUE], MolLogP=[VALUE],
HBD=[VALUE], HBA=[VALUE],
violations=[VALUE]

Molecule:
SMILES: [QUERY_SMILES]

Reply with EXACTLY ONE WORD:
either 'Yes' or 'No'.
Output nothing else. No explanation,
no punctuation, no context.
```

Prompt 5: Structure Retrieval Only

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

Structurally similar training molecules:
Example 1: SMILES=[TRAIN_SMILES_1],
label=[LABEL_1], sim=[SIM_1]
...
Example 5: SMILES=[TRAIN_SMILES_5],
label=[LABEL_5], sim=[SIM_5]

Query molecule:
SMILES: [QUERY_SMILES]

Reply with EXACTLY ONE WORD:
either 'Yes' or 'No'.
Output nothing else. No explanation,
no punctuation, no context.
```

Prompt 6: Full MOLE-RAG Classification

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

Retrieved chemistry context:
[TOP-5_RETRIEVED_PASSAGES]

Structurally similar training molecules:
Example 1: SMILES=[TRAIN_SMILES_1],
label=[LABEL_1], sim=[SIM_1]
...
Example 5: SMILES=[TRAIN_SMILES_5],
```

```
label=[LABEL_5], sim=[SIM_5]

Compound identifiers:
Name: [COMMON_NAME]
Other names: [SYNONYMS]
IUPAC: [IUPAC_NAME]

Functional groups:
[FUNCTIONAL_GROUPS]

RDKit descriptors:
[DESCRIPTOR_1]: [VALUE_1]
...
[DESCRIPTOR_15]: [VALUE_15]

Lipinski Rule of 5:
MolWt=[VALUE], MolLogP=[VALUE],
HBD=[VALUE], HBA=[VALUE],
violations=[VALUE]

Now predict the answer for this molecule.
SMILES: [QUERY_SMILES]

Reply with EXACTLY ONE WORD:
either 'Yes' or 'No'.
Output nothing else.
```

Prompt 7: Full MOLE-RAG Regression

```
System:
You are a chemistry expert.

User:
Task: Predict [TASK_DESCRIPTION].

Retrieved chemistry context:
[TOP-5_RETRIEVED_PASSAGES]

Structurally similar training molecules:
Example 1: SMILES=[TRAIN_SMILES_1],
true value=[VALUE_1], sim=[SIM_1]
...
Example 5: SMILES=[TRAIN_SMILES_5],
true value=[VALUE_5], sim=[SIM_5]

Compound identifiers:
Name: [COMMON_NAME]
Other names: [SYNONYMS]
IUPAC: [IUPAC_NAME]

Functional groups:
[FUNCTIONAL_GROUPS]

RDKit descriptors:
[DESCRIPTOR_1]: [VALUE_1]
...
[DESCRIPTOR_15]: [VALUE_15]

Target SMILES: [QUERY_SMILES]

Predict the [REGRESSION_PROPERTY]
for THIS target molecule.
Valid predictions are real numbers
[VALID_RANGE].

CRITICAL: Descriptor values and
molecular identifiers are properties of
the molecule, NOT the answer. Do not
echo those numbers.

Output EXACTLY ONE number
(example: [EXAMPLE_VALUE]).
No units, no descriptor names,
no explanation.

ANSWER:
```

Prompt 8: Synonym Cache Filtering

```
System:
You output JSON only. No prose,
no markdown.

User:
You are filtering chemical compound
synonyms for a literature search.

Keep paper-friendly names, such as
drug names, trade names, trivial names,
and IUPAC-style descriptive names.

Drop catalog, registry, database,
supplier, CAS, and lab notebook codes,
such as NSC IDs, CHEMBL IDs, NCI60 IDs,
SCHEMBL IDs, AKOS IDs, and CAS numbers.

Return only the paper-friendly names.
Order them by how commonly they appear
in chemistry literature.

IUPAC name: [IUPAC_NAME]
Candidate synonyms: [CANDIDATE_SYNONYMS]

Respond with JSON only:
{"keep": []}

Use an empty list if none are
paper-friendly.
```