GAMIC: Graph-Aligned Molecular In-context Learning for Molecule Analysis via LLMs

Ali Al-Lawati[1], **Jason Lucas**[1], **Zhiwei Zhang**[1], **Prasenjit Mitra**[2], **Suhang Wang**[1]
[1]The Pennsylvania State University, [2]Carnegie Mellon University Africa
{aha112,jsl5710,zbz5349,szw494}@psu.edu, prasenjit@cmu.edu

Abstract

In-context learning (ICL) effectively conditions large language models (LLMs) for molecular tasks, such as property prediction and molecule captioning, by embedding carefully selected demonstration examples into the input prompt. This approach eliminates the computational overhead of extensive pre-training and fine-tuning. However, current prompt retrieval methods for molecular tasks rely on molecule feature similarity, such as Morgan fingerprints, which do not adequately capture the global molecular and atom-binding relationships. As a result, these methods fail to represent the full complexity of molecular structures during inference. Moreover, medium-sized LLMs, which offer simpler deployment requirements in specialized systems, have remained largely unexplored in the molecular ICL literature. To address these gaps, we propose a self-supervised learning technique, GAMIC (Graph-Aligned Molecular In-Context learning), which aligns global molecular structures, represented by graph neural networks (GNNs), with textual captions (descriptions) while leveraging local feature similarity through Morgan fingerprints. In addition, we introduce a Maximum Marginal Relevance (MMR) based diversity heuristic during retrieval to optimize input prompt demonstration samples. Our experimental findings using diverse benchmark datasets show GAMIC outperforms simple Morgan-based ICL retrieval methods across all tasks by up to 45%.[1]

1 Introduction

Molecular representation and analysis field has significantly advanced towards specialized pre-trained language models like ChemBERTa (Chithrananda et al., 2020), and MolT5 (Edwards et al., 2022). Through targeted pre-training and task-specific fine-tuning, researchers have achieved state-of-the-art (SOTA) results in molecular property prediction (Tong et al., 2022; Liu et al., 2023a), molecule captioning (He et al., 2024; Jiang et al., 2024), and yield prediction (Guo et al., 2023; Shi et al., 2024).

Nonetheless, recent developments in large language models (LLMs) have demonstrated remarkable capabilities in prediction tasks through in-context learning (ICL) (Brown et al., 2020), potentially offering a more efficient alternative to the computationally expensive pre-train and fine-tune paradigm. Generally, in molecule captioning or property prediction tasks using LLMs, ICL retrieves molecules with similar captions or properties for a given molecule and uses these retrieved examples as in-context demonstrations (Guo et al., 2023; Li et al., 2024a). These demonstrations provide crucial guidance to help the LLM make more accurate predictions. While this approach can improve prediction accuracy, its effectiveness largely depends on the relevance and diversity of the selected examples (Ye et al., 2023). However, the effectiveness of ICL remains underexplored in molecular tasks, particularly for medium-sized LLMs ($<$ 10B) (Wang et al., 2024a), such as Mistral-7B (Jiang et al., 2023).

Recently, researchers have introduced Morgan fingerprint-based methods, such as *Scaffold* (Lim et al., 2020), for ICL demonstration selection, which utilizes the similarity of the Morgan fingerprint between the test sample and the demonstration pool (Guo et al., 2023). Although Scaffold outperforms random selection, its reliance on Morgan fingerprints only constrains its ability to retrieve structurally similar samples for ICL, as Morgan fingerprints cannot fully encode the complex binding relationships that are better represented by molecular graphs (Jin et al., 2018). Thus, capturing the graph structure is crucial for molecular analysis because it preserves atoms' spatial and connectivity information. This detailed representation is partic-

[1]Our code is available at: https://github.com/aliwister/mol-icl

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18364–18378
November 4-9, 2025

ularly important for molecular similarity retrieval, where subtle structural variations can significantly impact chemical behavior. This raises a natural question: **Can we combine the graph representation of the molecule with the Morgan fingerprint to further enhance ICL effectiveness by capturing both local properties (captured in the Morgan fingerprint) and global molecular structures (represented by a graph)?**

To explore this possibility, a leading approach is to leverage Graph Neural Networks (GNNs) (Scarselli et al., 2008), which are the SOTA method for processing molecular graphs (Wang et al., 2022b). However, applying GNNs in molecular similarity retrieval presents several challenges. In particular, (i) GNN encoding struggles to convert complex discrete molecular structures into continuous latent spaces while preserving chemical validity (Edwards et al., 2021), i.e. *complexity challenge*; (ii) GNN learning on multimodal datasets, such as PubChem (Kim et al., 2019), is susceptible to information loss due to the significant gap between graph and text representations (Song et al., 2024), i.e., *modality gap*; (iii) public datasets describe molecules in various ways, ranging from concise single-sentence descriptions to detailed multi-sentence explanations that capture very specific details (Liu et al., 2023b), i.e., *dataset limitations*, which further exacerbates the modality gap.

To address these challenges, we propose GAMIC (**G**raph-**A**ligned **M**olecular **I**n-**C**ontext learning), a novel ICL method that leverages the inherent graph structure of molecules and their local molecular features for multimodal graph-language training. In particular, GAMIC processes the molecular representation using a hierarchical graph encoder and aligns the latent representation with their scientifically-aware (e.g., SciBERT (Beltagy et al., 2019)) embedded textual descriptions using a sampling method based on Morgan fingerprint similarity. Incorporating Morgan fingerprints as a preliminary step in selecting alignment pairs helps narrow the *modality gap* by providing a robust and interpretable measure of local molecular similarity during multimodal alignment training. In addition, using scientifically-aware textual embedding enriches the latent space representation of the encoded graph post-alignment, mitigating the *complexity challenge*. Finally, having multiple potential textual representations for a molecule provides a more robust solution to address *dataset limitations* and mitigate inherent differences in the way captions describe molecules (for example, some molecules are described in multiple sentences and others in a few words).

Moreover, to further enhance ICL retrieval, we introduce a novel diversity-aware sample selection method using Maximum Marginal Relevance (MMR) to optimize the information provided in the input prompt.

Our **main contributions** are: (i) we propose a novel multimodal ICL framework for molecular tasks using graph molecular features grounded on Morgan fingerprint-based sampling; (ii) we propose an MMR-based demonstration selection heuristic to enhance sample diversity; and (iii) we conduct comprehensive experimental evaluations that demonstrates the effectiveness of our framework.

2 Related Work

2.1 Molecular Representation Learning

Traditional molecular modeling approaches have predominantly relied on specialized architectures that directly operate on molecular structures for tasks such as property prediction (Guo et al., 2021; Stärk et al., 2022), molecule generation (Gong et al., 2024; Kim et al., 2024), and reaction prediction (Liu et al., 2024). With the advent of the transformer architecture (Vaswani, 2017), the field has witnessed a shift towards representation learning through pre-training and fine-tuning paradigms. Early transformer-based approaches focused on learning from SMILES (Weininger, 1988) string representations. For example, MolBERT (Li and Jiang, 2021) adapted the BERT (Devlin et al., 2019) architecture to recognize different SMILES string representations of compounds, while ChemBERTa (Chithrananda et al., 2020) employed masked language modeling (MLM) on text-SMILES datasets. More recent approaches have explored richer molecular representations and transfer learning. MolT5 (Edwards et al., 2022) fine-tuned a pre-trained T5 language model for molecular translation. MolCA (Liu et al., 2023b) introduced a cross-model projector to effectively fine-tune LLMs on downstream tasks, while 3D-MolM (Li et al., 2024b) enhanced existing datasets by incorporating 3D conformational information generated using GPT-3.5.

Despite their effectiveness in molecular representation learning and analysis, these pre-train/fine-tune approaches encounter the following limita-

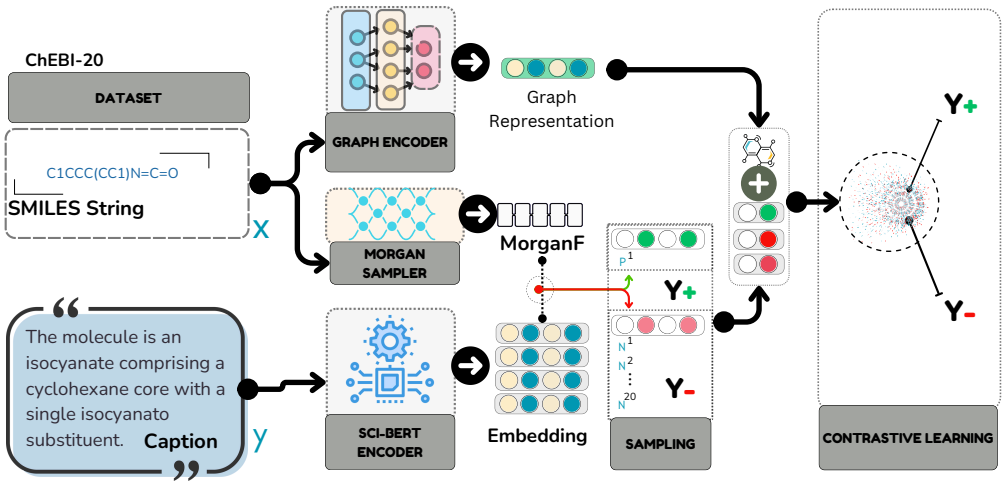

Figure 1: Overview of GAMIC Graph Projector

tions: (a) requiring significant computational resources during pre-training, (b) necessitating task-specific fine-tuning for each task, and (c) limited flexibility in adapting to new molecular tasks.

2.2 In-Context Learning for Molecular Tasks

ICL has emerged as a promising alternative to the traditional pre-train/fine-tune paradigm, enabling general-purpose language models to perform a variety of tasks through demonstration-based prompting. Rather than updating model weights, ICL conditions the model on task-specific demonstrations provided in the prompt, which guides it to generate more accurate responses. Despite the effectiveness of ICL in various applications (Dong et al., 2022; Al Lawati et al., 2025), its usage in molecular tasks is still in its early stage and there are few works (Guo et al., 2023; Li et al., 2024a) exploring this direction. Guo et al. (2023) establish a benchmark across eight molecular tasks, evaluating various LLMs using random and scaffold-based sample selection. MoleReGPT (Li et al., 2024a) similarly utilized scaffold-based retrieval for molecule captioning, but proposed fine-tuning for other tasks. Despite their effectiveness, existing ICL approaches for molecular tasks have several limitations: (a) insufficient capture of bond connectivity and atomic features present in molecular graphs, (b) limited consideration of the semantic richness enabled by text-informed graph modeling, and (c) overemphasis on large and commercial models, such as GPT-4.

While GNNs have shown promise in capturing molecular structure in fine-tuned model such as MolCA (Liu et al., 2023b), their potential for enhancing ICL demonstration selection remains underexplored. Our work addresses this gap by introducing GAMIC, the first approach to leverage Morgan-based graph alignment for ICL, which achieves SOTA performance on benchmark molecular ICL tasks. This novel direction addresses the limitations of existing methods while exploiting the computational efficiency central to the ICL paradigm.

3 Methodology

In this section, we first present the problem definition, then provide an overview of the proposed GAMIC, followed by a detailed description of its components.

3.1 Problem Setup

Given a training set $\mathcal{T} = (x_i, y_i)_{i=0}^{n}$ of molecule-value pairs with x_i as a SMILES string and y_i as the corresponding value, we aim to learn a GAMIC retriever, R, such that given a test molecule x_t, R can retrieve relevant and diverse demonstration $P_t = R(x_t, \mathcal{T})$ from a demonstration pool. P_t is concatenated with x_t to construct an LLM prompt for molecular analysis. The objective of the GAMIC retriever is to select P_t, such that $\mathcal{M}(P_t; x_t)$ will yield y'_t, that maximizes $\mathcal{D}(y_t, y'_t)$, where $\mathcal{D}$ is a similarity metric (e.g., BLEU score (Papineni et al., 2002)), ';' represents concatenation, and $\mathcal{M}$ represents the inference LLM.

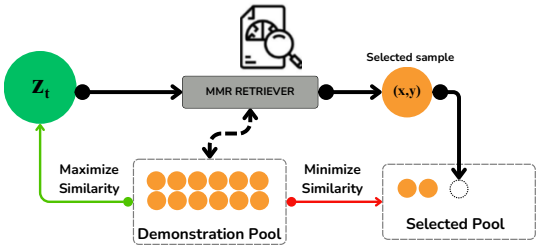

Figure 2: MMR-based sample selection

3.2 Overview of Model Architecture

Our proposed framework, GAMIC, is composed of two parts, i.e., (i) Graph Projection (see Figure 1), which aims to learn graph representation of a text-informed molecular graph that captures both bond connectivity and atomic features for demonstration retrieval; and (ii) MMR-based sample selection (see Figure 2), which aims to select similar and diverse demonstrations to improve the performance of an LLM.

Specifically, the graph projection adopts a **Graph Encoder** to learn the representation of molecular graphs constructed from SMILES strings. To train the graph encoder, we adopt contrastive learning and utilize a **Morgan Sampler** to find positive and negative candidates for alignment. The encoder is trained to learn the graph representation that aligns with positive textual captions encoded using the **SciBERT Encoder** using **Contrastive Learning**, as depicted in Figure 1

During the ICL demonstration retrieval process, a **MMR-based Sample Selector** retrieves informative and diverse examples. Next, we describe each component of GAMIC in more detail.

3.3 Graph Projection

Graph projection captures the underlying molecular structure by effectively aligning the molecular representation with the textual representation as detailed below.

3.3.1 Graph Encoder

To sufficiently capture the bond connectivity and atomic features present in molecular graphs, given a training set of (x, y) pairs, where x is the SMILES string, and y is the natural language description, i.e. caption, we construct a molecular graph for each SMILES string (x): $G = (\mathbb{V}, \mathbb{E})$ with atoms as nodes $\mathbb{V} = \{v_1, \ldots, v_N\}$ and bonds as edges $\mathbb{E}$. With the molecular graph, we apply a Graph Attention Network (GAT) (Veličković et al., 2017), which effectively captures the heterophily inherent in molecular graphs (Gao et al., 2023) through attention-based neighborhood aggregation (e.g., by appropriately weighting dissimilar neighbors). We adopt a two-layer GAT, a standard configuration that offers sufficient learning flexibility while avoiding over-smoothing. The node representations are learned as:

$$\mathbf{H} = \text{GAT}(\mathbf{X}, \mathbf{A}, \mathbf{E}; \theta_{GAT}), \quad (1)$$

where $\mathbf{A}$, $\mathbf{X}$, and $\mathbf{E}$ are the adjacency matrix, node features, and edge features, respectively. Next, we apply a pooling of node representations followed by an MLP to obtain the final graph embedding, $\mathbf{z}$, as follows:

$$\mathbf{z} = \text{MLP}(\text{MeanPool}\left(\mathbf{H}\right), \mathbf{w}^{(0)}), \quad (2)$$

where $\mathbf{w}^{(0)}$, is a learnable weight, and σ is the ReLU activation.

3.3.2 Morgan Sampler

In order to train the graph projector to align the final graph embedding with the captions, we propose adopting contrastive learning. Our preliminary testing showed that multimodal contrastive learning significantly outperforms other graph-based approaches such as graph autoencoder, or traditional graph-based contrastive methods. Hence, for each graph, we treat the corresponding caption as positive, and randomly sample negative pairs from the dataset. However, this may cause information loss due to the modality gap, as discussed above. In addition, dataset limitations, characterized by varying number of sentences in the captions or the level of details provided in the dataset, may hinder a robust alignment.

To address these issues, we propose Morgan fingerprint-based sampling ($\mathcal{R}_m$) to expand positive caption pairs according to high Morgan fingerprint similarity, while negative pairs are sampled based on low similarity. For each training sample, x_i, $\mathcal{R}_m(x_i)$ returns $\mathcal{Y}_i^+$, a set of positive samples and $\mathcal{Y}_i^-$, a set of negative samples, according to Morgan fingerprint similarity between x_i and the training set at each epoch.

3.3.3 SciBERT Encoder

To align the graph embeddings with texts, we need to encode the textual captions first. We adopt SciBERT (Beltagy et al., 2019) as the text encoder.

Task	Task Class	Dataset	Test Size	ICL Pool Size	Ev. Metrics
Molecule captioning	Molecular explaining	ChEBI-20	3300	26407	BLEU, ROUGE, METEOR
		PubChem	2000	12000	
Yield prediction	Molecular reasoning	Suzuki-Miyaura	576	4608	F1-score/StDev
		Buchwald-Hartwig	396	3163	
Property prediction	Molecular understanding	BBBP	204	1631	F1-score/StDev
		BACE	152	1209	
		HIV	4113	32901	
		Tox21	784	1184	
		ClinTox	148	6264	

Table 1: Overview of tasks, datasets, and evaluation metrics (Ev. Metrics)

SciBERT is a domain-specific model trained on a large corpus of scientific texts, providing better coverage of scientific terminology in molecular captions compared to general-purpose models (Li et al., 2024b), such as BERT. In our setting, we use SciBERT directly, without any task-specific fine-tuning. Specifically, for each caption $y \in \{\mathcal{Y}^+, \mathcal{Y}^-\}$, we obtain a fixed-size embedding using SciBERT as: $y_{emb} = \text{SciBERT}(y)$.

3.3.4 Contrastive Learning

Existing work on ICL has been limited by a lack of focus on graph-aware contrastive learning. To address this limitation, we propose utilizing a contrastive loss (Oord et al., 2018) that aligns the graph embeddings with their corresponding textual embeddings. The contrastive loss is formulated as:

$$\mathcal{L} = \text{NCE}(\mathbf{z}, \mathcal{Y}^+_{emb}, \mathcal{Y}^-_{emb}) \tag{3}$$

The Noise Contrastive Estimation (NCE) function is defined as:

$$\text{NCE}(\mathbf{z}, \mathcal{Y}^+, \mathcal{Y}^-) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\mathbf{z}_i \cdot y_i^+/\tau)}{\exp(\mathbf{z}_i \cdot y_i^+/\tau) + \sum_{j=1}^{K}\exp(\mathbf{z}_i \cdot y_{ij}^-/\tau)}\right) \tag{4}$$

where τ is a temperature parameter that controls the sharpness of the similarity distribution. The subscript $(_{emb})$ is omitted for all y in Equation (4) for readability.

3.4 MMR-based Sample Selector

During retrieval, we ensure both relevance and diversity in demonstration selection by employing a Maximal Marginal Relevance (MMR)-based selection strategy. For a given test sample (x_t, y_t), we select k demonstrations $(x_1, y_1), \ldots, (x_k, y_k)$ by iteratively optimizing:

$$\min_{\mathbf{z}\in P} \|\mathbf{z}_i - \mathbf{z}_t\| + \lambda \sum_{j=1}^{i-1} \max \|\mathbf{z}_i - \mathbf{z}_j\| \quad \text{for} \quad i \in 1, \ldots, k \tag{5}$$

where P is the set of possible demonstrations, $\mathbf{z}$ is the latent representation of x, and λ is a hyperparameter that balances relevance to the test sample (minimizing $\|\mathbf{z}_i - \mathbf{z}_t\|$) with diversity among the selected demonstrations (maximizing $\|\mathbf{z}_i - \mathbf{z}_j\|$). This approach ensures that selected demonstrations are both closely related to the test sample and diverse enough to improve the model's robustness. The selected demonstrations are appended in the prompt in reverse order, which improves prediction compared to other permutations (Lu et al., 2022).

4 Experiments

In this section, we conduct experiments to verify the effectiveness of our proposed framework. In particular, our aim is to answer the following research questions: (**RQ1**) *Molecular Performance Analysis*: How does the performance of ICL with `GAMIC` compare to other ICL methods on molecular analysis tasks? (**RQ2**) *Sensitivity Analysis*: How sensitive is `GAMIC` w.r.t to the number of demonstrations? (**RQ3**) *Ablation Study*: How does each `GAMIC` component contribute to the framework?

4.1 Experiment Setup

We evaluate our approach on three representative molecular tasks: molecule captioning, molecule property prediction, and molecule yield prediction, which represent three different molecular task classes, as summarized in Table 1.

4.1.1 Datasets

For each task, we utilize two or more datasets as follows:

- **Molecule captioning**: We evaluate molecule captioning using ChEBI-20 (Edwards et al., 2021) and PubChem (Kim et al., 2019) datasets. ChEBI-20 provides a focused assessment of bidirectional translation between molecular struc-

tures and natural language descriptions. *We also utilize the training set of this dataset to train* GAMIC. For PubChem, we utilize the test split proposed by Liu et al. (Liu et al., 2022).

- **Property prediction**: Datasets *BBBP*, *BACE*, *HIV*, *Tox21*, and *ClinTox* (Wu et al., 2018) are binary classification benchmarks containing SMILES strings and associated molecular property labels, which we use to evaluate prediction accuracy.
- **Yield prediction**: We utilize Suzuki-Miyaura (Reizman et al., 2016), and Buchwald-Hartwig (Ahneman et al., 2018) datasets which include molecule reactions and their corresponding yields, which can be classified as high or low.

For datasets without a predefined test split, we create three random train-validation-test splits using an 8:1:1 ratio, following standard practice (Wang et al., 2022a) using predefined random seeds. We conduct experiments on each split and report the average results across the three runs. Table 1 summarizes the key statistics of each dataset.

4.1.2 Baselines Molecular ICL Methods

As our framework focuses on ICL, we compare GAMIC with representative and state-of-the-art ICL methods for molecular analysis, including: (1) **Random**, which selects samples for the demonstration pool at random without replacement; (2) **Scaffold** (Guo et al., 2023), which utilizes Tanimoto similarity (Bajusz et al., 2015) between the Morgan fingerprints of the test sample and the demonstration pool to return the top k demonstrations, and (3) **GAE**, which utilizes a graph autoencoder (Kipf and Welling, 2016) to learn molecular graph representations and guide the selection of demonstration samples. Specifically, GAE adopts a two-layer GAT followed by a pooling layer to obtain a molecular graph embedding, then reconstructs the adjacency matrix with an MLP. The model is trained with MSE loss between the original and reconstructed adjacency matrices. Post training, the encoder utilizes latent structure to retrieve similar molecules.

4.1.3 LLM Models

To show that our GAMIC is flexible to facilitate various LLM backbones, we conduct comprehensive evaluations using three representative medium-sized LLMs, selected for their diversity in architecture and training approaches, which include (1) **Mistral-7B** (Jiang et al., 2023): a state-of-the-art model with 7 billion parameters, showcasing cutting-edge performance; (2) **OpenChat-8B** (Wang et al., 2024b): an open-source conversational model trained using high-quality dialogues to achieve performance on par with larger proprietary models; (3) **Zephyr-7B** (Tunstall et al., 2024): a fine-tuned variant of the Mistral architecture, optimized for specialized tasks.

Since this domain is highly specialized, our preliminary testing has shown that certain popular mediums-sized LLMs perform poorly on ICL molecular inference tasks, irrespective of the retrieval strategy. Consequently, we have tested multiple LLMs and prioritized the above mentioned models that have demonstrated more robust and consistent performance. Additional results for **Qwen-2.5-7B** (Qwen et al., 2025) and **Meta-Llama-3-8B** (Grattafiori et al., 2024) are reported in Appendix C.

4.1.4 Evaluation Metrics

For property prediction and yield prediction, we report the F1-score and the standard deviation. For molecule captioning, we utilize a comprehensive set of text generation metrics used in the literature (Guo et al., 2023; Li et al., 2024a) to evaluate molecular description quality: BLEU (BLEU-2, and BLEU-4), ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), and METEOR. All metrics range from 0 to 1, with higher scores indicating better alignment between generated and reference molecular descriptions.

4.1.5 Evaluation Setup

For each task, we follow the benchmark's standard evaluation protocol by evaluating the test set, and utilizing the training set as a demonstration pool from which samples can be retrieved, as described in Table 1.

To account for the stochastic nature of LLM outputs, we perform five repeated evaluations for each experiment and report the mean of the results. We evaluate our proposed method on the 9 different benchmark datasets across three molecular tasks.

For molecule captioning, we use $k = 2$ to control the prompt length as the labels for this task are long textual descriptions. For other tasks, we use $k = 3$. In addition, for all experiments, we set $\lambda = 0.3$ (in Equation (5)).

Dataset	Model	Method	Results					
			BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
ChEBI-20	Mistral	Random	0.229	0.125	0.325	0.152	0.273	0.287
		Scaffold	0.380	0.281	0.447	0.288	0.391	0.396
		GAE	0.492	0.386	0.574	0.414	0.515	0.536
		GAMIC	**0.542**	**0.439**	**0.617**	**0.466**	**0.561**	**0.585**
	OpenChat	Random	0.218	0.119	0.331	0.158	0.276	0.263
		Scaffold	0.363	0.269	0.446	0.286	0.391	0.381
		GAE	0.477	0.375	0.569	0.410	0.511	0.522
		GAMIC	**0.527**	**0.427**	**0.612**	**0.462**	**0.558**	**0.571**
	Zephyr	Random	0.177	0.093	0.304	0.139	0.258	0.252
		Scaffold	0.369	0.271	0.446	0.283	0.390	0.397
		GAE	0.477	0.372	0.561	0.401	0.503	0.521
		GAMIC	**0.526**	**0.422**	**0.605**	**0.451**	**0.548**	**0.570**
PubChem	Mistral	Random	0.155	0.084	0.251	0.122	0.215	0.210
		Scaffold	0.261	0.182	0.371	0.229	0.323	0.343
		GAE	0.318	0.242	0.437	0.299	0.390	0.403
		GAMIC	**0.340**	**0.262**	**0.455**	**0.317**	**0.407**	**0.421**
	OpenChat	Random	0.128	0.067	0.251	0.119	0.212	0.215
		Scaffold	0.203	0.140	0.360	0.221	0.313	0.336
		GAE	0.302	0.226	0.428	0.289	0.381	0.395
		GAMIC	**0.311**	**0.236**	**0.443**	**0.305**	**0.396**	**0.413**
	Zephyr	Random	0.149	0.080	0.250	0.121	0.214	0.206
		Scaffold	0.262	0.180	0.367	0.220	0.316	0.326
		GAE	0.310	0.235	0.427	0.291	0.382	0.392
		GAMIC	**0.323**	**0.246**	**0.441**	**0.304**	**0.394**	**0.406**

Table 2: Molecule captioning results using different ICL retrieval methods

4.2 RQ1. Molecular Performance Analysis

Molecule Explaining. Table 2 presents the results of GAMIC compared to benchmark methods on ChEBI-20 and PubChem datasets. GAMIC significantly outperforms other models across all evaluation metrics. This validates that graph representations capture the complex relationships of molecules more accurately. Furthermore, this demonstrated the effectiveness of GAMIC in mitigating the modality gap and dataset limitations present in both datasets.

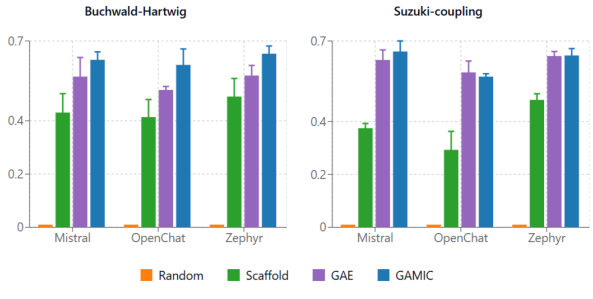

Figure 3: Yield prediction F1-score

Molecular Reasoning. As Figure 3 shows, GAMIC significantly improves the accuracy of yield prediction across all dataset and LLM combinations, which demonstrates it's ability to address the GNN complexity challenge. Hence, chemical validity is preserved in yield prediction more effectively than other baseline methods. Moreover, Random performs extremely poorly on both datasets on this task. On the other hand, GAE outperforms Scaffold, which validates the importance of graphs in accurately representing molecules.

Molecular Understanding. Table 4 shows the results for molecular understanding. GAMIC provides the best overall results on average, while Scaffold outperforms Random. On the HIV dataset using Random, Mistral reports an F1-score of 0, indicating its failure to identify any true positives. Overall, GAMIC outperforms the baselines on all property prediction benchmarks. The effectiveness of GAMIC on this task further corroborates its capacity to preserve chemical validity in cross-modal training.

Method	Results					
	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Random	0.229	0.125	0.325	0.152	0.273	0.287
Scaffold	0.380	0.281	0.447	0.288	0.391	0.396
GAE	0.492	0.386	0.574	0.414	0.515	0.536
GAMIC	**0.542**	**0.439**	**0.617**	**0.466**	**0.561**	**0.585**

Table 3: Molecule captioning results of Qwen-2.5-32B (Qwen et al., 2025) on ChEBI-20 dataset using GAMIC and baseline ICL retrieval methods

Applicability to Larger LLMs. While our focus has been on medium-sized LLMs that are associated with lower computational costs and easier

Model	Method	BBBP	BACE	HIV	Tox21	ClinTox	All Data Mean
Mistral	Random	0.694 ± 0.032	0.372 ±0.062	0	0.037 ± 0.025	0.011 ± 0.043	0.223
	Scaffold	0.850 ± 0.494	0.710 ± 0.093	0.392 ± 0.216	0.203 ± 0.099	0.100 ± 0.087	0.451
	GAE	0.858 ± 0.012	0.701 ± 0.053	0.289 ± 0.012	0.216 ± 0.068	0.103 ± 0.178	0.433
	GAMIC	**0.905** ± 0.031	**0.726** ± 0.127	**0.400** ± 0.202	**0.271** ± 0.064	**0.112** ± 0.040	**0.483**
OpenChat	Random	0.289 ± 0.051	0.525 ± 0.005	0.012 ± 0.013	0.008 ± 0.013	0.044 ± 0.077	0.176
	Scaffold	0.749 ± 0.022	0.665 ± 0.053	0.364 ± 0.018	0.111 ± 0.085	0.083 ± 0.144	0.394
	GAE	0.745 ± 0.013	0.674 ± 0.021	0.315 ± 0.055	0.131 ± 0.059	0.048 ± 0.082	0.383
	GAMIC	**0.836** ± 0.024	**0.674** ± 0.037	**0.365** ± 0.019	**0.153** ± 0.019	**0.203** ± 0.093	**0.446**
Zephyr	Random	0.518 ± 0.034	0.750 ± 0.032	0.020 ± 0.009	0.095 ± 0.040	0.139 ± 0.127	0.304
	Scaffold	0.875 ± 0.004	0.769 ± 0.040	0.386 ± 0.054	0.242 ± 0.046	0.242 ± 0.162	0.503
	GAE	0.881 ± 0.022	0.747 ± 0.065	0.326 ± 0.037	0.246 ± 0.021	0.169 ± 0.177	0.474
	GAMIC	**0.924** ± 0.009	**0.783** ± 0.034	**0.422** ± 0.011	**0.276** ± 0.023	**0.361** ± 0.127	**0.553**

Table 4: Property prediction F1-score and a summarized mean score

deployment in real-world applications, we briefly explore GAMIC's performance on larger LLMs by evaluating it on Qwen-2.5-32B, as reported in Table 3.

The results suggest the contributions of GAMIC extend to larger LLMs and improve their performance significantly compared to baseline ICL retrieval methods.

4.3 RQ2: Sensitivity Analysis

Model	k	Results					
		BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Mistral	0	0.055	0.023	0.135	0.065	0.123	0.073
	1	0.536	0.431	0.612	0.459	0.554	0.581
	2	0.542	0.439	0.617	0.466	0.561	0.585
	3	**0.543**	**0.440**	**0.619**	**0.468**	**0.563**	**0.586**
	4	0.531	0.426	0.609	0.454	0.551	0.573
	5	0.530	0.425	0.609	0.454	0.551	0.573
	10	0.528	0.423	0.605	0.450	0.547	0.572
OpenChat	0	0.037	0.007	0.101	0.011	0.083	0.067
	1	0.523	0.422	0.606	0.455	0.550	0.569
	2	0.527	0.427	0.613	0.462	0.557	0.571
	3	**0.528**	0.427	**0.614**	**0.461**	**0.557**	**0.573**
	4	0.518	0.416	0.603	0.449	0.547	0.563
	5	0.521	0.419	0.609	0.456	0.553	0.569
	10	0.518	0.415	0.605	0.449	0.549	0.563
Zephyr	0	0.048	0.005	0.130	0.018	0.100	0.082
	1	0.514	0.409	0.592	0.438	0.535	0.558
	2	0.526	0.422	0.605	0.451	0.548	0.570
	3	0.526	**0.423**	**0.609**	**0.455**	**0.552**	0.570
	4	0.524	0.419	0.606	0.451	0.549	0.568
	5	0.520	0.416	0.605	0.449	0.547	0.565
	10	0.518	0.412	0.599	0.442	0.540	0.563

Table 5: Sensitivity analysis for different ICL demonstration sample sizes (k) on molecule captioning

We conduct a sensitivity analysis to assess how molecule captioning performs in response to additional demonstration samples. Specifically, we vary the number of demonstrations as $\{0, 1, 2, 3, 5, 10\}$. As reported in Table 5, the results plateau at three ICL samples and there is insignificant improvement between $k = 2$, and $k = 3$, which further motivates our selection of $k = 2$ for this task to control prompt length. As we increase $k > 3$, the performance begins to deteriorate slowly.

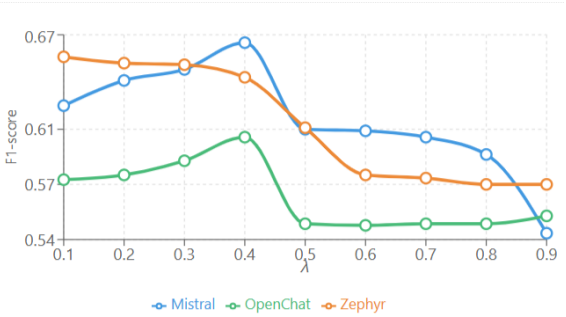

Figure 4: λ sensitivity analysis using average yield prediction

Furthermore, we analyze the sensitivity of the MMR parameter, λ, on the prediction outcome. We fix $k = 3$ and vary λ from 0.1 to 0.9. Based on the results in Figure 4, we observe that $\lambda = 0.3$ or $\lambda = 0.4$ are plausible configurations.

4.4 RQ3: Ablation Study

We conduct a focused ablation study to evaluate the contribution of each module to our framework by comparing GAMIC against the following variants: (i) W/o Morgan-BERT: During training, this method uses only the corresponding caption as the positive pair, and other samples as negative pairs. It also encodes captions with BERT, which has limited scientific vocabulary, rather than SciBERT. This helps isolate the contributions of SciBERT and Morgan sampling; (ii) GAMIC-BERT: Uses Morgan sampling during training, but encodes captions with BERT instead of SciBERT; (iii) W/o Morgan: Similar to (i), but encodes captions using SciBERT, which helps to quantify the individual contribution of SciBERT to GAMIC without Morgan sampling.

Model	Method	Morgan S	SciBERT	Results					
				BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Mistral	W/o Morgan-BERT	✗	✗	0.520	0.415	0.599	0.444	0.541	0.566
	GAMIC-BERT	✓	✗	0.533	0.430	0.611	0.457	0.553	0.577
	W/o Morgan	✗	✓	0.535	0.431	0.613	0.460	0.554	0.580
	GAMIC	✓	✓	**0.542**	**0.439**	**0.617**	**0.466**	**0.561**	**0.585**
OpenChat	W/o Morgan-BERT	✗	✗	0.505	0.404	0.594	0.441	0.538	0.551
	GAMIC-BERT	✓	✗	0.518	0.418	0.604	0.452	0.548	0.562
	W/o Morgan	✗	✓	0.522	0.421	0.608	0.456	0.552	0.566
	GAMIC	✓	✓	**0.527**	**0.427**	**0.613**	**0.462**	**0.557**	**0.571**
Zephyr	W/o Morgan-BERT	✗	✗	0.508	0.404	0.589	0.434	0.532	0.553
	GAMIC-BERT	✓	✗	0.520	0.416	0.600	0.445	0.543	0.565
	W/o Morgan	✗	✓	0.521	0.416	0.602	0.447	0.545	0.567
	GAMIC	✓	✓	**0.526**	**0.422**	**0.605**	**0.451**	**0.548**	**0.570**

Table 6: GAMIC ablation results on molecule captioning using ChEBI-20 dataset

Table 6 demonstrates the contribution of Morgan sampling and SciBERT compared to W/o Morgan-BERT. Both approaches contribute similarly on their own, with a slight advantage for using SciBERT. The combined contribution of both elements leads to better overall performance than either component individually. Additional comparisons with other BERT-based encoders are provided in Appendix D.

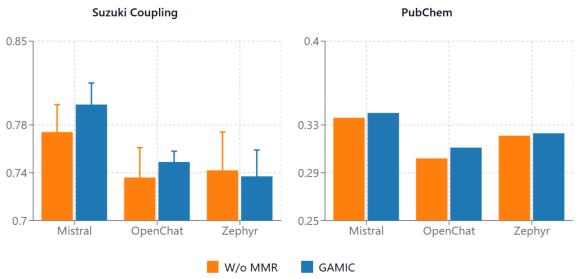

Figure 5: MMR vs Top-K on Suzuki dataset accuracy (left) and PubChem BLEU score (right)

Additionally, we evaluate the contribution of MMR by comparing it with Top-K, which retrieves top k most similar demonstrations ordered in reverse similarity.

Figure 5 illustrates the improvement of MMR in yield and property prediction averages. It shows that MMR provides better overall results across multiple molecular tasks and LLMs.

Figure 6 illustrates the selection strategies of Top-K and MMR in 2-D projected latent space. There is more overlap among the demonstrations selected by Top-K, whereas MMR produces greater diversity. Furthermore, while some demonstrations are common to both methods, the order in which these examples are selected differs.

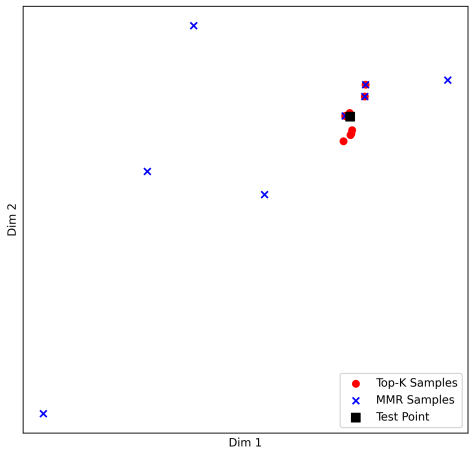

Figure 6: MMR vs Top-K demonstration selection on projected latent spaces

5 Conclusion

In this work, we demonstrate the potential of ICL in improving LLM performance on molecular tasks. We focus on medium-sized LLMs (7–10B parameters) due to their lower computational costs and ease of deployment in real-world applications.

While previous work has considered either Morgan similarity (scaffold) or graph-based similarity in isolation, we present GAMIC, which effectively combines the merits of both approaches. Morgan fingerprints encode molecular substructures, while graphs capture the complex interactions between individual elements. Combining both approaches provides a more holistic representation of the molecule. Our experiments demonstrate that our method achieves SOTA on 26 of 27 tests performed across three molecular tasks.

Acknowledgment

This material is based upon work supported by, or in part by the Army Research Office (ARO) under grant number W911NF-21-10198, the Department of Homeland Security (DHS) under grant number 17STCIN00001-05-00, and Cisco Faculty Research Award. The views and conclusions contained in this material are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

Limitations

We focus on medium-sized LLMs with lower computational costs and ease of deployment in real-world applications (<10B). We have not extensively examined the applicability of our approach to larger or proprietary models, although preliminary experiments with a larger model suggest that our method may generalize.

In addition, we have not considered the impact of using decoder-based LLMs instead of SciBERT (an encoder-based language model). Existing studies (Muennighoff et al., 2024) find that LLMs (decoder-only transformers) are good at generation, but not at representation learning, and will result in degraded performance. On the other hand, encoder-only transformers such as SciBERT are well-suited for representation learning. There has been recent work suggesting certain layers of LLMs may perform better at representation learning, however, the reported gains are marginal (Skean et al., 2025).

References

Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. 2018. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190.

Ali Al Lawati, Jason Lucas, and Prasenjit Mitra. 2025. Semantic captioning: Benchmark dataset and graph-aware few-shot in-context learning for SQL2Text. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8026–8042, Abu Dhabi, UAE. Association for Computational Linguistics.

Dávid Bajusz, Anita Rácz, and Károly Héberger. 2015. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. volume 33, pages 1877–1901.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, and 1 others. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In *EMNLP*.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *EMNLP*.

Chen Gao, Yu Zheng, Nian Li, Yinfeng Li, Yingrong Qin, Jinghua Piao, Yuhan Quan, Jianxin Chang, Depeng Jin, Xiangnan He, and Yong Li. 2023. A survey of graph neural networks for recommender systems: Challenges, methods, and directions. *ACM Trans. Recomm. Syst.*, 1(1).

Haisong Gong, Qiang Liu, Shu Wu, and Liang Wang. 2024. Text-guided molecule generation with diffusion language model. In *AAAI*, volume 38, pages 109–117.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng

Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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

Zhichun Guo, Chuxu Zhang, Wenhao Yu, John Herr, Olaf Wiest, Meng Jiang, and Nitesh V Chawla. 2021. Few-shot graph learning for molecular property prediction. In *Proceedings of the web conference 2021*, pages 2559–2567.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh V. Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. *arXiv preprint*. ArXiv:2402.07630 [cs].

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*.

Yinuo Jiang, Xiang Zhuang, Keyan Ding, Qiang Zhang, and Huajun Chen. 2024. Enhancing cross text-molecule learning by self-augmentation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9551–9565.

Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakkola. 2018. Learning multimodal graph-to-graph translation for molecular optimization. *arXiv preprint arXiv:1812.01070*.

Seojin Kim, Jaehyun Nam, Sihyun Yu, Younghoon Shin, and Jinwoo Shin. 2024. Data-efficient molecular generation with hierarchical textual inversion. *arXiv preprint arXiv:2405.02845*.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, and 1 others. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109.

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

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*.

Juncai Li and Xiaofei Jiang. 2021. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021(1):7181815.

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

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. 2020. Scaffold-based molecular design with a graph generative model. *Chemical science*, 11(4):1153–1164.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shengchao Liu, Weili Nie, Chengpeng Wang, and 1 others. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *arXiv preprint arXiv:2212.10789*.

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

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024. Reactxt: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, SU Hongjin, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2024. text-embedding-3-large. https://platform.openai.com/docs/guides/embeddings. Embedding model released by OpenAI.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Brandon J Reizman, Yi-Ming Wang, Stephen L Buchwald, and Klavs F Jensen. 2016. Suzuki–miyaura cross-coupling optimization enabled by automated feedback. *Reaction chemistry & engineering*, 1(6):658–666.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Runhan Shi, Gufeng Yu, Xiaohong Huo, and Yang Yang. 2024. Prediction of chemical reaction yields with large-scale multi-view pre-training. *Journal of Cheminformatics*, 16(1):22.

Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. *arXiv preprint arXiv:2502.02013*.

Jia Song, Wanru Zhuang, Yujie Lin, Liang Zhang, Chunyan Li, Jinsong Su, Song He, and Xiaochen Bo. 2024. Towards cross-modal text-molecule retrieval with better modality alignment. *arXiv preprint arXiv:2410.23715*.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Liò. 2022. 3d infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning*, pages 20479–20502. PMLR.

Xiaochu Tong, Dingyan Wang, Xiaoyu Ding, Xiaoqin Tan, Qun Ren, Geng Chen, Yu Rong, Tingyang Xu, Junzhou Huang, Hualiang Jiang, and 1 others. 2022. Blood–brain barrier penetration prediction enhanced by uncertainty estimation. *Journal of Cheminformatics*, 14(1):44.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2024. Zephyr: Direct distillation of lm alignment. *CoLM*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, and 1 others. 2024a. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024b. Openchat: Advancing open-source language models with mixed-quality data. *ICLR*.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D Burke. 2022a. Chemical-reaction-aware molecule representation learning. *ICLR*.

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

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

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

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023. Complementary explanations for effective in-context learning. *Preprint*, arXiv:2211.13892.

A Prompt for zero-shot Molecule Captioning

For zero-shot molecular captioning experiments, we utilize the following prompt:

Zero-shot Prompt

You are an expert chemist. Given the molecular SMILES, your task is to predict the molecule description using your experienced molecular knowledge.

SMILES:[SMILES String]
Caption:

Model	Method	Results					
		BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Qwen-2.5-7B	GAMIC	0.496	0.401	0.607	0.460	0.552	0.577
Meta-Llama-3-8B		0.319	0.247	0.519	0.371	0.461	0.493

Table 7: GAMIC results using other medium-sized LLMs

Model	Encoder	Results					
		BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Mistral	BioBERT	0.538	0.434	0.615	0.463	0.557	0.581
	PubMedBERT	0.541	0.438	0.619	0.467	0.561	0.584
	SciBERT	0.542	0.439	0.617	0.466	0.561	0.585
	text-embedding-3-large	**0.550**	**0.445**	**0.625**	**0.472**	**0.566**	**0.591**
OpenChat	BioBERT	0.522	0.421	0.609	0.457	0.552	0.568
	PubMedBERT	0.526	0.425	0.613	0.460	0.557	0.569
	SciBERT	0.527	0.427	0.613	0.462	0.557	0.571
	text-embedding-3-large	**0.534**	**0.430**	**0.614**	**0.459**	0.554	**0.577**
Zephyr	BioBERT	0.525	0.420	0.607	0.451	0.548	0.570
	PubMedBERT	0.526	0.422	0.607	0.452	0.549	0.569
	SciBERT	0.526	0.422	0.605	0.451	0.548	0.570
	text-embedding-3-large	**0.535**	**0.434**	**0.619**	**0.468**	**0.561**	**0.570**

Table 8: GAMIC ablation results with different BERT configurations, and a large proprietary encoder. Best and second-best results for each model are highlighted in bold and underlined, respectively.

For multi-shot, we do not provide instructions. Instead, we begin the prompt with the ICL demonstrations in input/output format:

2-shot Prompt Captioning

SMILES: [Sample 1 SMILES]
Caption: [Sample 1 CAPTION]

SMILES: [Sample 2 SMILES]
Caption: [Sample 2 CAPTION]

SMILES: [Test SMILES]
Caption:

Depending on the dataset, the label was changed from 'Caption' to 'High-Yield' for yield prediction. For datasets: BBBP, BACE, HIV, Tox21, ClinTox, the label was changed to 'BBBP', 'BACE', 'HIV Active','NR-ER', and 'CT_TOX', respectively. This is consistent with the conventions used by Guo et al. (2023).

B Additional Data on Evaluation Metrics

For molecular explanation we utilize the following metrics:

- **BLEU** (Bilingual Evaluation Understudy) (Papineni et al., 2002): We use BLEU-2 and BLEU-4 scores to assess n-gram precision between generated and reference texts. BLEU-2 captures local phrase matching, while BLEU-4 evaluates longer sequence accuracy.

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004): We utilize three variants: (1) ROUGE-1: measures unigram overlap (2) ROUGE-2: assesses bigram overlap (3) ROUGE-L: evaluates longest common subsequence, capturing flexible sequence matching

- **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee and Lavie, 2005): provides a more nuanced evaluation by incorporating synonyms, stemming, and word order, to better capture semantic similarity.

C Additional LLMs

We performed preliminary tests on additional medium LLMs to evaluate their performance in molecular tasks, which motivated our choice of LLMs selected for the main experiments.

Table 7 reports molecule captioning results of GAMIC on Qwen-2.5-7B and Meta-Llama-3-8B. These results may be compared with Table 2.

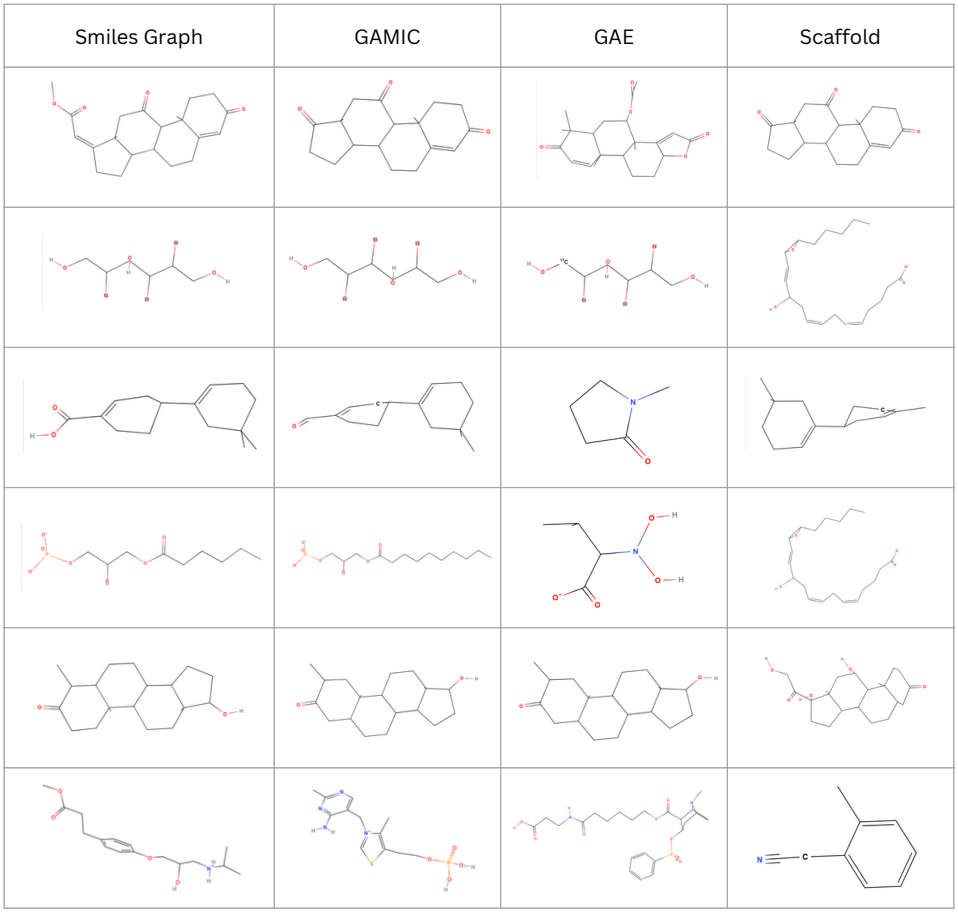

Figure 7: Retrieval examples using various methods

D Additional Embedding Models

In this section, we compare the performance of SciBERT with two domain-specific BERT variants: BioBERT (Lee et al., 2020) and PubMedBERT (Gu et al., 2021), as well as a proprietary large embedding model.

Table 8 shows the three BERT methods provide comparable results, with a slight edge for SciBERT. This may be due to SciBERT's broader semantic grounding and exposure to diverse scientific formats (Beltagy et al., 2019), making it better suited for graph-text alignment tasks such as GAMIC, despite being trained on less data than PubMedBERT.

Conversely, we observe that proprietary large-scale models, notably OpenAI's text-embedding-3-large (OpenAI, 2024), can outperform scientifically aware models, although their adoption will result in additional costs.

E MMR Time Complexity Analysis

Here, we establish an upper bound on the time complexity of MMR sample selection.

Let n denote the size of the demonstration pool. For a given test sample x_t, we first retrieve the top K nearest neighbors, where $K \ll n$ (typically, $K = 30$–50), using a priority queue-based selection in $O(n + K \log n)$ time.

Next, we apply the MMR selection procedure (as described in Equation (5)) to choose the final k demonstrations from the K candidates. This step requires:

$$O(K) + O(2K) + \cdots + O(kK) \leq O(Kk^2)$$

time, as each of the k selections involves scanning up to K candidates against previously selected items.

Hence, the time complexity of MMR-based demonstration retrieval is bounded by:

$$O(n + K \log n + Kk^2)$$

F Case Study

Figure 7 shows the molecules recovered for a set of test molecules using GAMIC alongside several baseline methods. It is visually apparent that GAMIC consistently retrieves molecular graphs that more closely resemble the global structure of the target molecule, including connectivity patterns. Conversely, the baselines often miss broader structural context. This highlights GAMIC's ability to capture and leverage global graph-level information more effectively during retrieval.

G Computational Experiments

All experiments were carried out using NVIDIA A100 40GB GPUs. Across all runs, the total computational cost amounted to approximately 310 GPU hours for the evaluations reported in this work. This estimate reflects efficient batching and optimization to minimize overhead, while ensuring reproducibility. This number does not include GPU usage for debugging, hyperparameter tuning, or other tests.