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

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 avoids the computational overhead of extensive pertaining and fine-tuning. However, current prompt retrieval methods for molecular tasks have relied 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, small-to-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%. Our code is available at https://anonymous.4open.science/r/mol-icl-AEF7.

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, given a target molecular for molecule captioning or property prediction using LLM, ICL retrieves similar molecules with their captions or properties, and uses these retrieved examples in the prompt as demonstration [Guo *et al.*, 2023; Li *et al.*, 2024a], which provides important information to guide LLMs to give more accurate predictions. While this approach can enhance prediction accuracy, its effectiveness heavily depends on both the relevance and diversity of the demonstration samples used to guide the LLM [Das *et al.*, 2021]. Despite this, the effectiveness of ICL remains underexplored in molecular tasks, particularly for small to medium-sized LLMs (< 10B) 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 [Guo *et al.*, 2023], which utilizes the similarity of the Morgan fingerprint between the test sample and the demonstration pool. 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 particularly 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 graph structures [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-Context 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 to select 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, by expanding the pool of potential textual representations grounded on Morgan fingerprints, GAMIC provides a more robust solution to address *dataset limitations*. Moreover, to further enhance ICL retrieval, we introduce a novel diversity-aware sample selection method using Maximum Marginal Relevance (MMR) to maximize the information provided in the input prompt.

Our key contributions are:

- A novel multimodal ICL method for molecular tasks using graph molecular features grounded on Morgan fingerprint-based sampling.
- An MMR-based demonstration selection heuristic to enhance sample diversity.
- Comprehensive experimental evaluation comparing our approach with existing methods using three medium-size general-purpose LLMs.

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] finetunes a pre-triend T5 language model for moleculecular translation. MolCA [Liu *et al.*, 2023b] introduced a cross-model projector to effectively fine-tune LLMs on select downstream tasks, while 3D MolM enhanced existing datasets by incorporating 3D conformational information generated using GPT-3.5.

Despite their effectiveness in molecular representation learning and analysis, these pre-training and fine-tuning approaches face the following limitations: (a) requirements for significant computational resources during pre-training, (b) need for task-specific fine-tuning and separate training 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 for the pre-train/fine-tune paradigm, enabling general-purpose language models to perform various tasks through demonstration-based prompting. Instead of fine-tuning, ICL provides demonstrations in the prompt, which allows the LLM to learn from them and generate more accurate responses. Despite the effectiveness of ICLs in various applications [Dong *et al.*, 2022], the work on ICL for molecular tasks is still in its early stage and there are very few works [Li *et al.*, 2024a; Guo *et al.*, 2023]. Recently, MoleReGPT [Li *et al.*, 2024a] introduced dual approaches for molecular tasks. For molecular captioning, MoleReGPT utilizes Morgan fingerprint similarity, i.e., Scaffold, which compares the presence of specific substructures encoded in the Morgan fingerprint vector. Guo et al. [Guo *et al.*, 2023] established a benchmark across eight molecular tasks, evaluating various LLMs using random and scaffold-based sample selection. However, existing ICL approaches for molecular tasks have several limitations: (a) insufficient capture of bond connectivity and atomic features present in molecular graphs, (b) limited exploration of graph-aware contrastive learning for demonstration selection and (c) primarily focus on large and commercial language models, such as GPT4.

While GNNs have demonstrated 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, achieving SOTA performance on benchmark molecular ICL tasks. This novel direction addresses the limitations of existing methods while maintaining computational efficiency central to the ICL paradigm.

3 Methodology

In this section, we will first give the problem definition, then the overview of the proposed GAMIC followed by the details of GAMIC.

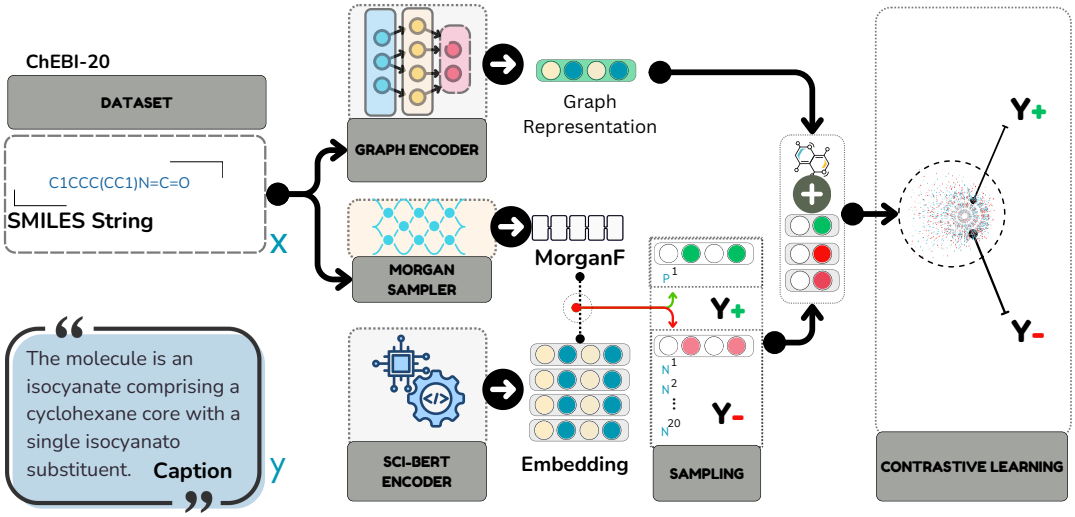

Figure 1: Overview of GAMIC Graph Projector

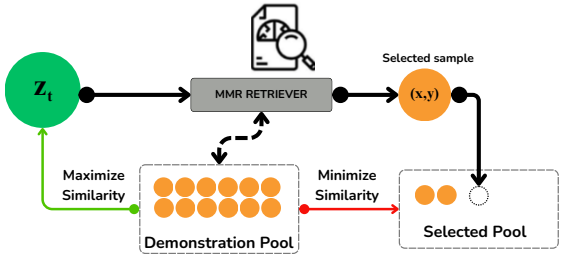

Figure 2: MMR-based Sample Selector

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 graph retriever R, such that given a test molecule x_t, the GAMIC retriever can retrieve relevant and diverse demonstration $P_t = R(x_t, \mathcal{T})$ from a demonstration pool, which will be concatenated with x_t and prompt as input to an LLM $\mathcal{M}$ 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]) and ';' represents concatenation.

3.2 Overview of Model Architecture

The proposed framework, GAMIC, is composed of two parts, i.e., (i) Graph Projection (see Figure 1), which aims to learn graph representation of a molecualr 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 graph representation of molecular graphs constructed from SMILE Strings. To train the graph encoder, it adopts contrastive learning and utilizes a **Morgan Sampler** to find positive and negative alignment candidates for contrastive learning. The encoder is trained to learn graph representation that align with positive textual captions encoded using the **SciBERT Encoder** using Contrastive Learning, as depicted in fig. 1 . During the ICL demonstration retrieval process, MMR-based Sample Selector retrieves informative and diverse examples. Next, we describe each component of GAMIC in more detail.

3.3 Graph Projection

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 use a two-layer Graph Attention Network (GAT) [Veličković *et al.*, 2017] to learn node representation 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 on top on node representation followed by a 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 weights, and σ is ReLU activation.

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

Table 1: Overview of tasks, datasets, and evaluation metrics

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	

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 a negative pair(s) 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 type of details described, may hinder a robust alignment.

Therefore, we propose adopting Morgan fingerprint-based sampling ($\mathcal{R}_m$) to expand the sets of positive and negative caption pairs for alignment by including molecules with similar Morgan fingerprints. 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, based on Morgan fingerprint similarity between x_i and the training set at each epoch.

SciBERT Encoder

To align the graph representation with texts, we need to get text representation first. we adopt SciBERT [Beltagy *et al.*, 2019] as the text encoder. 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] like BERT. Specifically, for each caption $y \in \{\mathcal{Y}^+, \mathcal{Y}^-\}$, we obtain a fixed-size embedding using SciBERT as:

$$y_{emb} = \text{SciBERT}(y) \tag{3}$$

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 graph embeddings with their corresponding text representations. The contrastive loss is formulated as:

$$\mathcal{L} = \text{NCE}(\mathbf{z}, \mathcal{Y}^+_{emb}, \mathcal{Y}^-_{emb}), \tag{4}$$

where the Noise Contrastive Estimation (NCE) function is defined as:

$$\text{NCE}(\mathbf{z}, \mathcal{Y}^+, \mathcal{Y}^-) = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp\left(\mathbf{z}_i \cdot y_i^+/\tau\right)}{\exp\left(\mathbf{z}_i \cdot y_i^+/\tau\right) + \sum_{j=1}^{K}\exp\left(\mathbf{z}_i \cdot y_{ij}^-/\tau\right)}\right) \tag{5}$$

where τ is a temperature parameter that controls the sharpness of the similarity distribution, and subscript $(_{emb})$ is omitted for all y 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 solving the following optimization iteratively:

$$\min_{\mathbf{z}\in P} \|\mathbf{z}_i - \mathbf{z}_t\| + \lambda \sum_{j=1}^{i-1} \max \|\mathbf{z}_i - \mathbf{z}_j\| \quad \text{for} \quad i \in 1, \ldots, k \tag{6}$$

where P is the set of possible demonstrations and $\mathbf{z}$ 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 as depicted in fig. 3, which improves prediction compared to other permutations [Lu *et al.*, 2022].

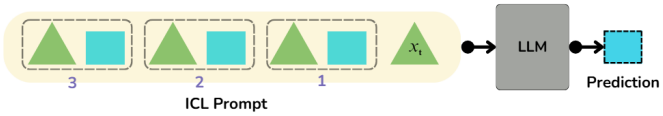

Figure 3: Triangles represent SMILES strings, and squares are the labels. The ICL samples are appended in reverse order of retrieval.

4 Experiments

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

4.1 Experiment Setup

Datasets

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 (See Table 1). For each task, we utilize two or more datasets as follows:

Table 2: Molecule captioning test results using different ICL retrieval methods

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**

- **Molecule Captioning**: We evaluate performance on molecule captioning using the test split of ChEBI-20 [Edwards *et al.*, 2021]. This dataset provides a focused assessment of bidirectional translation between molecular structures and natural language descriptions. *We also utilize the training set of this dataset to train* GAMIC. In addition, we utilize the split suggested by Liu et al. [Liu *et al.*, 2022] to evaluate the PubChem [Kim *et al.*, 2019] dataset.
- **Property Prediction**: Datasets *BBBP*, *BACE*, *HIV*, *Tox21*, and *ClinTox* proposed by [Wu *et al.*, 2018] are binary classification datasets that consist of SMILES strings, and binary labels of specific molecular properties, which we use to assess the accuracy of the predictions.
- **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 yield which can be classified as high or low.

For datasets without a predefined test split, we create three random train-valid-test splits by 8:1:1 ratio, following standard practice in the literature [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 the datasets.

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 Selection**, 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 demonstrations to return the top k demonstrations. The demonstrations are appended in reverse order as in fig. 3; and (3) **GAE**, which utilizes graph autoencoder [Kipf and Welling, 2016] to learn graph representations. Specifically, it adopts a two-layer GAT followed by a pooling layer to obtain graph representation for a molecular graph, then reconstruct the adjacency matrix with an MLP and adopts mean square loss between the original adjacency matrix and the reconstructed adjacency matrix as the loss function to train the autoencoder. Once the model is trained, the encoder can utilize latent structure for retrieving similar molecules.

LLM Models

To show that our GAMIC is flexible to facilitate various LLM backbones, we conduct comprehensive evaluations using three representative small to medium-sized Language Models (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.*, 2024]: An open-source conversational AI model, highlighting the strengths of publicly accessible systems; (3) **Zephyr-7B** [Tunstall *et al.*, 2024]: A fine-tuned variant of the Mistral architecture, optimized for specialized tasks.

Evaluation Metrics

For property prediction and yield prediction, we report the F1-score and the standard deviation. For molecule captioning, we employ 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.

Table 3: Property Prediction F1-score and a summarized mean score

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**

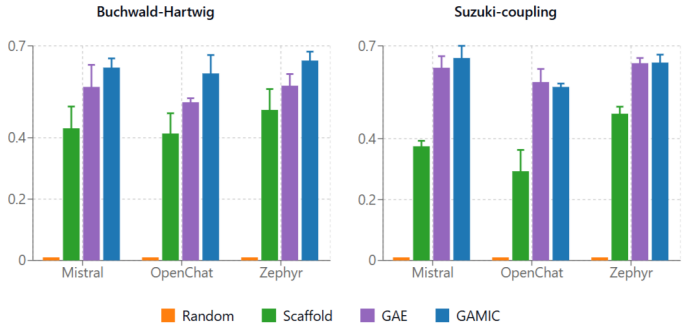

Figure 4: Yield prediction F1-score

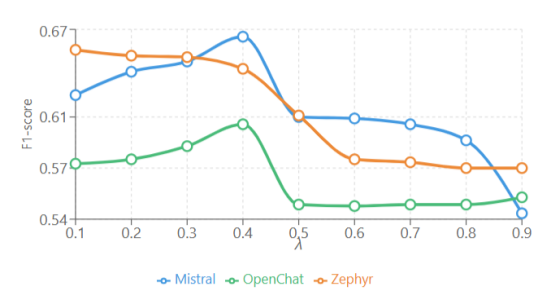

Figure 5: λ sensitivity analysis using average Yield prediction

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 use $\lambda = 0.3$.

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 effectively. Furthermore, this demonstrated the effectiveness of GAMIC in overcoming the modality gap and dataset limitations present in both datasets.

Molecular Reasoning. As fig. 4 shows, GAMIC significantly improves the accuracy of yield prediction across all dataset/LLM combinations, which demonstrates it's effectiveness in overcoming the GNN complexity challenge. Hence, chemical validity is preserved in yield prediction more effectively than other baseline methods.

Moreover, random selection performs extremely poorly on both datasets on this task. On the other hand, GAE outperforms Scaffold, which validates the importance of graphs in effectively representing molecules.

Molecular Understanding. Table 3 shows the results for molecular understanding. GAMIC provides the best overall results on average, while Scaffold outperforms random selection. On the HIV dataset using Random retrieval, Mistral reports an F1-score of 0, indicating a failure to achieve 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.

4.3 RQ2: Sensitivity Analysis

We conduct a sensitivity analysis to assess how the molecule captioning performs in response to additional demonstration samples. Specifically, we vary the number of demonstrations as $\{0, 1, 2, 3, 5, 10\}$ and the results are given 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.

Furthermore, we analyze how modifying the MMR parameter, λ, affects the prediction outcome. We fix k as 3 and vary λ from 0.1 to 0.9. The results are shown in Figure 5. Based

Table 4: GAMIC ablation results on molecule captioning using ChEBI-20 dataset

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

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

on the figure, we can observe that values of 0.3 or 0.4 appear plausible choices.

4.4 RQ3: Ablation Study

We conduct a focused ablation study to evaluate the contribution of each module to our framework by comparing it 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 a 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 to quantify the contribution of SciBERT.

Table 4 demonstrates the contribution of Morgan sampling and SciBERT compared to W/o Morgan-BERT. Both approaches contribute similarly to individual improvements, with a slight advantage for using SciBERT. The combined contribution of both elements leads to better performance than either method alone.

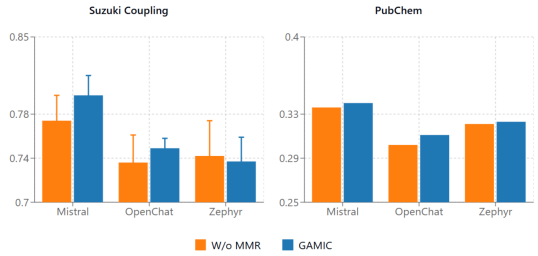

Figure 6: MMR vs W/o MMR on Suzuki dataset accuracy (left) and PubChem BLEU score (right)

Additionally, we evaluate the contribution of MMR by comparing it with W/o MMR, which retrieves the top k most similar samples, ordered in reverse similarity, as shown in Figure 2. Figure 6 illustrates the improvement of MMR in yield and property prediction averages. It shows that MMR provides better results across multiple tasks and for all LLMs tested.

5 Conclusions

This work demonstrates the potential of medium-sized Large Language Models (LLMs) in molecular understanding. We focus on smaller LLMs (7–10B parameters) due to their lower computational costs and ease of deployment in real-world applications. Our results demonstrate the capacity of these LLMs to perform multiple molecular tasks without task-specific fine-tuning using advanced demonstration selection techniques. We introduced GAMIC, which achieves state-of-the-art performance in molecular ICL. These findings bridge the gap between molecular structure representation and LLM capabilities, advancing applications in drug discovery and materials science.

Ethical Statement

There are no ethical issues.

Acknowledgments

References

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

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

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

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

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

[Das *et al.*, 2021] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases. In *EMNLP*, November 2021.

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

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

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

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

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

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

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

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

[Jiang *et al.*, 2023] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

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

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

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

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

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

[Li *et al.*, 2024a] Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

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

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

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

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

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

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

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

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

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002.

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

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

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

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

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

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

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

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

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

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

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

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

[Weininger, 1988] David Weininger. 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, 1988.

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