M^2LLM: Multi-view Molecular Representation Learning with Large Language Models

Jiaxin Ju[1], **Yizhen Zheng**[2], **Huan Yee Koh**[2,3], **Can Wang**[1] and **Shirui Pan**[1†]
[1]School of Information and Communication Technology, Griffith University
[2]Department of Data Science and AI, Monash University
[3]Drug Discovery Biology, Monash Institute of Pharmaceutical Sciences, Monash University
jiaxin.ju@griffithuni.edu.au, {yizhen.zheng,huan.koh}@monash.edu, {can.wang,s.pan}@griffith.edu.au

Abstract

Accurate molecular property prediction is a critical challenge with wide-ranging applications in chemistry, materials science, and drug discovery. Molecular representation methods, including fingerprints and graph neural networks (GNNs), achieve state-of-the-art results by effectively deriving features from molecular structures. However, these methods often overlook decades of accumulated semantic and contextual knowledge. Recent advancements in large language models (LLMs) demonstrate remarkable reasoning abilities and prior knowledge across scientific domains, leading us to hypothesize that LLMs can generate rich molecular representations when guided to reason in multiple perspectives. To address these gaps, we propose M^2LLM, a multi-view framework that integrates three perspectives: the molecular structure view, the molecular task view, and the molecular rules view. These views are fused dynamically to adapt to task requirements, and experiments demonstrate that M^2LLM achieves state-of-the-art performance on multiple benchmarks across classification and regression tasks. Moreover, we demonstrate that representation derived from LLM achieves exceptional performance by leveraging two core functionalities: the generation of molecular embeddings through their encoding capabilities and the curation of molecular features through advanced reasoning processes.

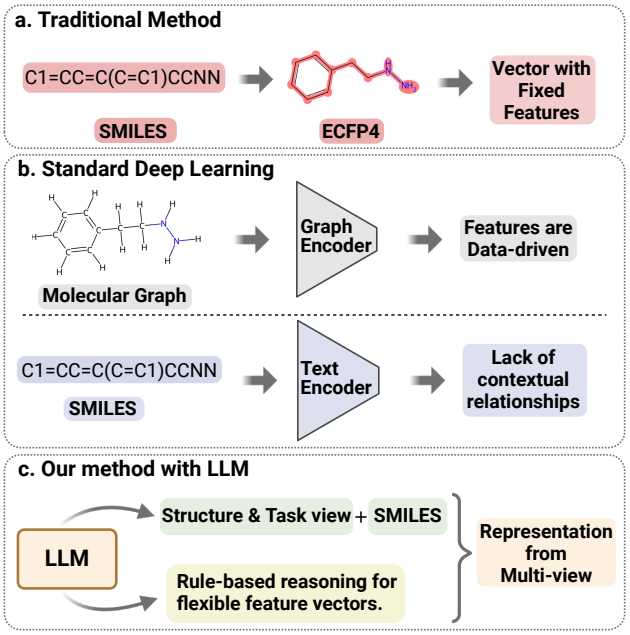

Figure 1: (a) Traditional Method: Converts SMILES into fixed ECFP4 vectors. (b) Standard Deep Learning: Graph encoders learn patterns from data, while text encoders for SMILES-only input lack contextual relationships. (c) M^2LLM: Generating multi-view representations by leveraging two core capabilities of LLMs: encoding contextual relationships and rule-based reasoning for molecular feature generation.

1 Introduction

Molecular property prediction is vital in cheminformatics [Yang *et al.*, 2019] and drug discovery [Drews, 2000], enabling the estimation of key characteristics like blood-brain barrier permeability, solubility, and toxicity. Traditional approaches rely heavily on predefined molecular descriptors, such as Extended-Connectivity Fingerprints (ECFPs) [Rogers and Hahn, 2010], derived from SMILES(Simplified Molecular Input Line Entry System) [Weininger, 1988]. While these methods are efficient, their fixed feature sets limit their ability to capture the complex relationships needed for specific chemical tasks. Graph Neural Networks (GNNs) have demonstrated strong capabilities across a wide range of domains [Bu *et al.*, 2024; Yu *et al.*, 2024; Wang *et al.*, 2024a; Wang *et al.*, 2024b] and have also shown effectiveness in capturing structural and physicochemical patterns from molecular graphs [Koh *et al.*, 2024; Yu *et al.*, 2025; Du *et al.*, 2024]. However, they rely heavily on dataset-driven learning, which limits their ability to generalize across diverse chemical tasks. Similarly, recent studies [Sadeghi *et al.*, 2024; Shirasuna *et al.*, 2024] leverage text encoder [Medsker *et al.*, 2001] or Transformer-based language model [Kenton and

† Corresponding author.

Toutanova, 2019] using SMILES representations, which focus solely on string patterns and lack the ability to capture contextual relationships or molecular semantics.

To address these limitations, large language models (LLMs) offer a transformative approach to molecular representation by leveraging semantic and contextual knowledge from diverse pretraining corpora, including scientific literature and domain-specific datasets [Zheng *et al.*, 2024c]. In addition, LLMs demonstrate emergent abilities such as contextual reasoning, relational understanding, and the ability to extrapolate patterns, making them uniquely suited for tasks that require deep semantic understanding [Kojima *et al.*, 2022; Zheng *et al.*, 2025]. These capabilities far surpass those of earlier language models or text encoders. While those methods demonstrate the utility of LLMs in capturing molecular representations, they are inherently constrained by their exclusive reliance on SMILES as input. As a result, the potential of LLMs to enhance molecular representation by leveraging their pretrained scientific knowledge and emergent abilities remains underexplored.

In this work, we propose M^2LLM, a novel multi-view molecular representation learning framework that addresses these gaps by fully exploiting the power of LLMs. The framework is organized into two key modules: Molecular Embedding Generation and Molecular Feature Curation, each contributing distinct yet complementary perspectives on molecular data. The molecular embedding generation module leverages the semantic embedding ability of LLMs to represent molecular information from multiple views. This module includes the molecular structure view, which encodes structural information from SMILES sequences to capture structure-specific insights. This structure view can be further extended to incorporate additional insights, broadening the scope and enriching the molecular representation. Additionally, the molecular task view contextualizes molecules within specific prediction tasks to provide task-relevant guidance. Together, these views harness the LLM's pretrained semantic knowledge to generate comprehensive embeddings.

The molecular feature curation module, on the other hand, utilizes the reasoning ability of LLMs to derive interpretable features. This module introduces the molecular rules view, which generates rule-based features informed by scientific knowledge and observed data patterns, facilitating a deeper understanding of molecular properties. To unify these representations, M^2LLM employs a dynamic fusion mechanism that adaptively combines the contributions of each view based on the requirements of the task and the characteristics of individual molecules. The fused representation is then used for downstream prediction tasks, with a multi-layer perceptron (MLP) designed for both classification and regression. Through extensive experiments, we demonstrate that M^2LLM achieves state-of-the-art performance across multiple molecular property prediction benchmarks, highlighting the potential of LLMs to redefine molecular representation learning.

Our contributions of this work are as follows: (1) We propose M^2LLM, a novel multi-view molecular representation learning framework that integrates diverse molecular perspectives through molecular structure view, molecular task view, and molecular rules view. These views are fused dynamically to create a unified representation tailored to each prediction task. (2) We explore the potential of representations derived from LLMs using M^2LLM for molecular property prediction, demonstrating that LLMs can achieve high performance by leveraging their dual capabilities: molecular embedding generation through their encoding abilities and molecular feature curation through their reasoning capabilities, addressing a significant gap in current research. (3) We show that M^2LLM achieves state-of-the-art performance across multiple molecular property prediction benchmarks, demonstrating its adaptability, scalability, and effectiveness in advancing the use of LLMs for molecular representation learning.

2 Related Work

Several molecular representation methods have been developed, including molecular graph, ECFPs, and string line annotations such as SMILES. With the advancement of AI, machine learning models are now extensively used for property prediction through traditional and deep learning approaches.

In the conventional approach, traditional machine learning models, such as random forests [Breiman, 2001], rely on computed molecular fingerprints to predict properties by capturing relationships between molecular substructures [Jeon and Kim, 2019], though these predefined fingerprints may not fully capture complex molecular structural patterns and interactions. On the other hand, GNNs have been widely applied across domains [Zheng *et al.*, 2024a; Wu *et al.*, 2024; Zhang *et al.*, 2025; Zhang *et al.*, 2019], they have also been effectively used to model molecular graphs [You *et al.*, 2020; Wang *et al.*, 2022; Xia *et al.*, 2022], capturing hierarchical structural information and uncovering complex molecular patterns. However, these approaches often fail to integrate broader contextual, and may overlook knowledge already discovered in scientific literature and encoded within LLMs.

LLMs like GPT-4 [OpenAI *et al.*, 2023] and Galactica [Taylor *et al.*, 2022], trained on diverse scientific and chemical datasets, capture semantic and contextual relationships beyond traditional text encoders' syntactic patterns. Moreover, LLMs exhibit emergent abilities such as reasoning [Kojima *et al.*, 2022], relational understanding [Mirza *et al.*, 2024], and pattern recognition [Zheng *et al.*, 2025], making them powerful tools for extracting insights from molecular text representations. Mirza et al. [2024] indicate that even a 7B-parameter LLM can achieve average human scores, while advanced models like GPT-4 can surpass the highest human scores in chemical reasoning.

Previous studies [Wang *et al.*, 2019; Fabian *et al.*, 2020; Ross *et al.*, 2022] have explored encoding SMILES using LLMs as molecular embeddings, demonstrating their effectiveness in capturing meaningful representations and performance in downstream property prediction tasks [Sadeghi *et al.*, 2024]. Researchers [Luo *et al.*, 2024; Zheng *et al.*, 2024b; Rollins *et al.*, 2024] have also investigated combining text encoders with GNNs to leverage both contextual and structural information. However, these methods remain limited by their reliance on SMILES as the sole input, failing to fully exploit the semantic depth of LLMs.

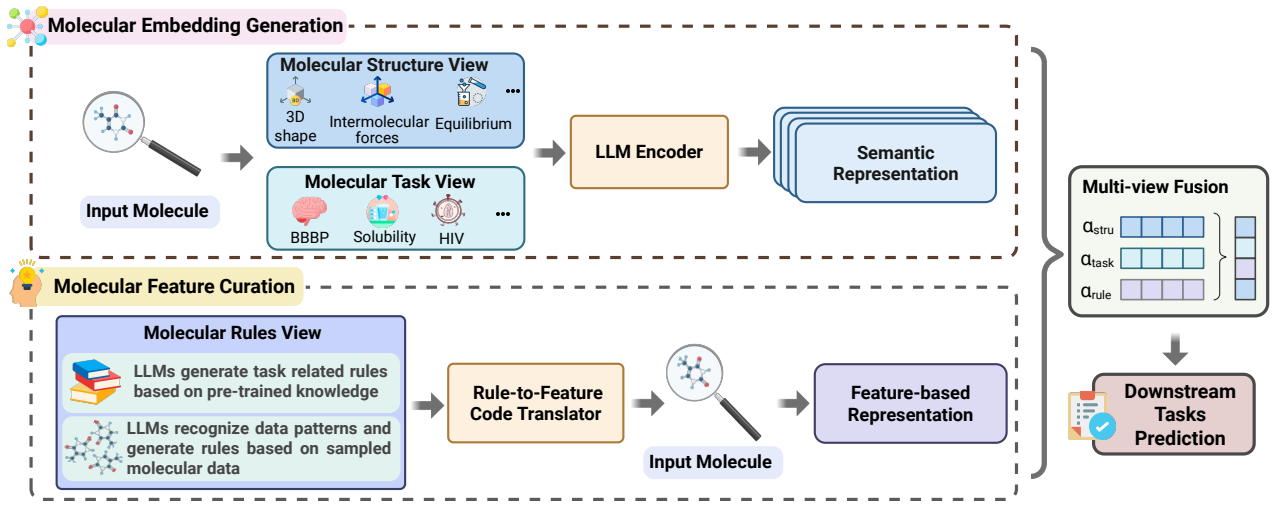

Figure 2: **The M^2LLM Molecular Representation Learning Framework.** The framework integrates three molecular views to create comprehensive and adaptable representations. Embedding generation includes two views: the molecular structure view and the molecular task view, both processed using LLMs to produce semantic representations. Feature curation provides one view: the molecular rules view, where rules are generated using LLMs based on pretrained knowledge and recognized data patterns, and translated into features via a rule-to-feature translator to produce feature-based representation. These diverse representations are fused in a multi-view fusion module, with trainable weights (α) balancing each component, enabling accurate predictions through a MLP prediction.

3 Multi-view Molecular Representation Learning with Large Language Models

In this section, we introduce M^2LLM, a novel multi-view framework designed to leverage LLMs for molecular representation learning. As illustrated in Figure 2, the framework consists of two main components: Embedding Generation and Feature Curation, which collectively provide three distinct views of molecular information. These views are fused in a Multi-View Fusion Module, generating a unified representation optimized for downstream prediction tasks.

3.1 Molecular Embedding Generation

The Molecular Embedding Generation component leverages LLMs to encode sequence inputs, providing two views: the Molecular Structure View and the Molecular Task View.

Molecular Structure View

The molecular structure view uses LLMs to generate embeddings that capture key physical and chemical properties of molecules. LLMs have demonstrated strong capabilities in encoding semantic and contextual knowledge. These models generate outputs sequentially, relying on previous input tokens to reason about the current context. Instead of directly providing SMILES as input, we frame specific questions about the molecule alongside its SMILES representation. This approach gives the LLM additional contextual information, allowing it to "think about" the molecule in relation to the queried property, resulting in richer and more meaningful representations. To achieve this, we define three example insights as questions targeting various structural aspects of the molecule, which can be adapted or replaced as needed.

> ***Structure Insight 1:*** How does the molecule's 3D shape change in different environments, and what are the effects of these changes?
> ***Structure Insight 2:*** What are the key intermolecular forces that govern the behavior of this molecule in various contexts?
> ***Structure Insight 3:*** How does the molecule contribute to the overall chemical equilibrium in its different environments?

Let s_i denote the SMILES for molecule i, and q_j denote one of the three structure insight questions. The structural embedding z_{ij}^{struct} for molecule i with question q_j is computed as:

$$z_{ij}^{\text{struct}} = f_{\text{Encode}}^{\text{struct}}([q_j; s_i]) \tag{1}$$

where $f_{\text{Encode}}(\cdot)$ is the LLM encoder used to process the input, $[q_j; x_i]$ represents the concatenation of the question q_j and the SMILES s_i. The embeddings generated for the three questions are then concatenated to form the final structure view representation:

$$z_i^{\text{struct}} = [z_{i1}^{\text{struct}}; z_{i2}^{\text{struct}}; z_{i3}^{\text{struct}}] \tag{2}$$

Molecular structure view not only enriches the molecular structure representation but also provides a flexible framework for generating diverse and comprehensive embeddings. By incorporating structural views, M^2LLM focuses solely on analyzing molecular information without being tied to specific tasks or datasets. This design allows the LLM to encode the input in a way that leverages the semantic knowledge learned during the pretraining process, capturing more contextual and meaningful information than using SMILES alone. Additionally, the modular nature of the structural

view facilitates straightforward extensions by introducing new questions to create additional views.

Molecular Task View

The molecular task view encodes task-specific information by framing molecular analysis as a natural language processing problem, inspired by the pre-training process of the Galactica model [Taylor *et al.*, 2022] that has demonstrated the ability to process and reason about molecular representations effectively when supplemented with contextual information, such as questions or prompts. To leverage this capability, the task view combines a molecule's SMILES representation with a task-specific question, guiding the LLM to generate a task-aware representation by retrieving and processing semantic information encoded during pretraining. For example, consider the task of predicting blood-brain barrier penetration for a given molecule. The input to the LLM would take the following form:

> **Here is a SMILES formula:** [START_I_SMILES]C1=CC=C(C=C1)C(=O)O[END_I_SMILES]
>
> **Question:** Will the chemical compound penetrate the blood-brain barrier?

Unlike the molecular structure view, which focuses on general analysis of molecular properties, the task view explicitly tailors its input to the prediction problem at hand. The generated embedding z_i^{task} incorporates both the SMILES sequence s_i and the task-specific question t, and is computed as follows:

$$z_i^{\text{task}} = f_{\text{Encode}}^{\text{task}}([t; s_i]) \tag{3}$$

3.2 Molecular Feature Curation

The Molecular Feature Curation component introduces an additional view of molecular representation through the Molecular Rules View. This view captures domain-specific knowledge and patterns by leveraging LLMs to generate rules based on pretrained knowledge and data-driven insights. These rules are then transformed into features using a Rule-to-Feature Code Translator, providing a numerical representation that complements the semantic representations generated by Molecular Embedding Generation module.

Molecular Rules View

The Molecular Rules View captures both pretrained knowledge from scientific literature and patterns derived from molecular datasets. The generated rules are transformed into numerical features, creating a representation that complements the semantic representations.

Scientific Rule Generation with LLMs: LLM has built-in knowledge and an understanding of various tasks from its pre-training. To make use of this, we assign the LLM a specific persona, such as an experienced chemist, and instruct it to generate rules based on its extensive exposure to scientific literature. These rules are generated independently of any specific molecule and are instead tailored to the requirements of a given task. By leveraging its pretraining on vast scientific datasets, the LLM produces rules that reflect well-established principles and patterns relevant to the task at hand. Let t represent the specific task, the generated rules, R_{sci}, for task t can be expressed as $R_{\text{sci}}(t) = f_{\text{Reason}}^{\text{sci}}(t)$, where $f_{\text{Reason}}^{\text{sci}}(\cdot)$ leverages the LLM's pretrained reasoning capabilities to derive rules. The example below demonstrates how the LLM generates scientifically grounded rules for predicting blood-brain barrier penetration in Scientific Rule Generation phase.

> **Persona**: Assume you are an experienced Chemist. Please come up with 20 rules that are important to predict if a molecule can penetrate the blood-brain barrier.
>
> **LLM Answer Example**:
> Rule 1: Molecular weight $<$ 500 Da
> Rule 2: LogP value between 1 and 3
> Rule 3: Presence of aromatic rings
> ...

Data Pattern Rule Observation with LLMs: Beyond their extensive knowledge base, LLMs exhibit strong abilities in identifying patterns and relationships [Zheng *et al.*, 2025], making them well-suited for recognizing task-relevant trends within molecular data. In this phase, several randomly selected subsets of SMILES strings in the training data $\{s_i\}_{i=1}^m \in S_{\text{train}}$ with their corresponding label y_i are then provided to the LLM. By analyzing these subsets, additional rules R_{data} based on observed patterns and relationships within the molecular structures for the specific task t: $R_{\text{data}}(t) = \bigcup_{k=1}^K f_{\text{Reason}}^{\text{data}}(\{s_i, y_i\}_{i=1}^m, t)_k$, where K is the number of subsets analyzed, m is the number of molecule in each subset, and $f_{\text{Reason}}^{\text{data}}(\cdot)$ leverages the LLM's emergent reasoning capabilities to identify patterns and generate rules. The following example illustrates how the LLM generates task-specific rules by analyzing randomly selected one subset of molecular training data paired with their corresponding labels.

> **Persona**: Assume you are a very experienced Chemist. In the following data, with label 1, it means the smiles string is BBBP. With label 0, it means the smiles string is not BBBP. Please infer step-by-step to come up with 3 rules that directly relate the properties/structures of a molecule to predict if it can be BBBP.
> [***SMILES strings along with corresponding labels***]
>
> **LLM Answer Example**:
> Rule 1: The presence of a benzene ring in the molecule is essential for predicting whether it can be BBBP.
> Rule 2: The presence of a carbonyl group (-C=O) in the molecule is also important for predicting whether it can be BBBP.
> Rule 3: Number of Hydrogen Bond Donors (HBD)

Feature-based Representation

The combined set of rules, $R(t) = R_{\text{sci}}(t) \bigcup R_{\text{data}}(t)$, is transformed into numerical features through a Rule-to-Feature Code Translator, which maps the rules,

$\{r_1, r_2, \ldots, r_n\} \in R(t)$, into a feature vector for each SMILES string s. This translation process leverages LLMs to transform text-based rules into code-based features and can be defined as:

$$f_i(s) = \begin{cases} 1, & \text{if } r_i \text{ is satisfied by } s \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

or alternatively as a numerical function $f_i(s) \in \mathbb{R}^n$ if the rule outputs a continuous value (e.g., molecular weight). The LLM is tasked with translating these rules into executable functions $f_i(\cdot)$, which are applied to the input s to extract a set of molecular features. This process generates a feature vector z_i^{rule} for a molecule s, where each feature corresponds to a rule:

$$z_i^{\text{rule}} = [f_1(s), f_2(s), \ldots, f_n(s)] \tag{5}$$

3.3 Multi-view Representation Fusion

The M^2LLM framework then integrates three views into a unified representation through a fusion mechanism. The fused representation is subsequently passed through a prediction module, which adapts to the requirements of either classification or regression tasks.

For a given molecule s_i, let z_i^{struct}, z_i^{task}, and z_i^{rule} denote the representations obtained from the structure view, task view, and rules view, respectively. To combine these representations into a single, comprehensive vector, our proposed framework employs a weighted sum mechanism. Each view's contribution is modulated by a set of learnable weights α_i^{struct}, α_i^{task}, and α_i^{rule}, which are specific to each molecule. The fused representation z_i^{fused} is then computed as a weighted sum of the individual view representations:

$$z_i^{\text{fused}} = \alpha_i^{\text{struct}} z_i^{\text{struct}} + \alpha_i^{\text{task}} z_i^{\text{task}} + \alpha_i^{\text{rule}} z_i^{\text{rule}} \tag{6}$$

where weights satisfy: $\alpha_i^{\text{struct}} + \alpha_i^{\text{task}} + \alpha_i^{\text{rule}} = 1$, $\alpha_i^{\text{struct}}, \alpha_i^{\text{task}}, \alpha_i^{\text{rule}} \geq 0$. The fused representation z_i^{fused} is then used as input to a multi-layer perceptron (MLP), which performs the final prediction. This process is mathematically defined as:

$$\hat{y}_i = f_{\text{MLP}}(z_i^{\text{fused}}) \tag{7}$$

where $\hat{y}_i$ is the predicted output for molecule s_i, and $f_{\text{MLP}}(\cdot)$ represents the function of the multi-layer perceptron.

The framework is trained to optimize the weights α_i^{struct}, α_i^{task}, and α_i^{rule}, and the MLP parameters using task-specific loss functions. For classification tasks, the cross-entropy loss is minimized:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] \tag{8}$$

For regression tasks, the root mean squared error (RMSE) loss is used:

$$\mathcal{L}_{\text{regression}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \tag{9}$$

where N denotes the number of molecules, y_i represents the binary ground truth label or the true value, and $\hat{y}_i$ is the predicted probability or the predicted value.

4 Experiment

4.1 Experimental Setup

Dataset Our framework is evaluated on 8 datasets spanning 34 tasks from MoleculeNet [Wu *et al.*, 2018], including physiology-related tasks like BBBP [Martins *et al.*, 2012], ClinTox [Gayvert *et al.*, 2016], and 27 SIDER tasks [Kuhn *et al.*, 2016] for adverse drug reaction prediction. Additionally, we evaluate classification tasks from BACE [Subramanian *et al.*, 2016] and HIV [Wu *et al.*, 2018], as well as regression tasks from ESOL [Delaney, 2004], FreeSolv [Mobley and Guthrie, 2014], and Lipophilicity [Wu *et al.*, 2018]. We use the scaffold splitting method recommended by MoleculeNet [Wu *et al.*, 2018], which assigns molecules with distinct structural scaffolds to separate training, validation, and test sets. This method, unlike random splitting, ensures structural dissimilarity between sets, creating a more challenging evaluation scenario. Detailed dataset descriptions are provided in Appendix A.1.

Baselines For the traditional model, we use Random Forest [Breiman, 2001] with ECFP4 [Rogers and Hahn, 2010] as the input feature set. For deep learning models, as shown in Table 1, we select the most representative GNNs pretraining baselines and transformer-based architecture. To ensure a fair comparison, We rerun all baseline models with the same random seed across 10 iterations.

Backbone Model Our framework is built on an LLM-based multi-view architecture, utilizing a state-of-the-art LLM as its backbone. Specifically, for the molecular structural and task views, we use Galactica models (6.7B and 30B parameters) [Taylor *et al.*, 2022], LLaMa-3.1 models (8B and 8B-instruct) [Dubey *et al.*, 2024], and OpenAI's closed-source text embedding models (small and large configurations) [OpenAI, 2024]. For the molecular rules view, rule generation is performed using the Galactica models, leveraging their extensive pretraining on scientific literature to produce high-quality task-specific rules.

4.2 Performance on Classification Tasks

We evaluate M^2LLM on five classification datasets with 31 subtasks, as shown in Table 1. We report the mean and standard deviation from 10 random seeds using the evaluation metric, receiver operating characteristic-area under the curve (ROC-AUC) (%), where higher scores indicate better performance. One result for the same LLM backbone architecture is presented for the comparison with other state-of-the-art baselines. Full results for different backbone architectures are provided in the Appendix A.2.

As shown in Table 1, our framework demonstrates superior performance, surpassing existing baselines with significant improvements. Notably, our framework exhibits exceptional performance on the Clintox dataset, achieving a near-perfect accuracy of 99.5% and 99.4%, this result significantly outperforms all other models. Moreover, on the BBBP, HIV, and SIDER datset, our framework variants achieve the best and second-best results, outperforming all GNN-based and Transformer-based baselines. This result further enhances the credibility of LLM-based approaches in molecular property prediction tasks. Full results on 27 tasks for the SIDER

	Model \ Dataset	Backbone Type	BBBP(1) ↑	BACE(1) ↑	ClinTox(1) ↑	HIV(1) ↑	SIDER(27) ↑	Average	Average Rank
	RF + ECFP4	RF	67.6 ± 1.0	**85.0 ± 1.2**	69.4 ± 3.1	77.1 ± 0.7	62.6 ± 2.5	72.3 ± 1.7	7
Baselines	**AttrMask** [Hu *et al.*, 2019]	GNN	65.2 ± 1.4	77.8 ± 1.8	73.5 ± 4.3	75.3 ± 1.5	55.7 ± 4.0	69.5 ± 2.6	10
Baselines	**GraphCL** [You *et al.*, 2020]	GNN	67.8 ± 2.4	74.6 ± 2.1	77.5 ± 3.4	75.1 ± 0.7	53.1 ± 4.3	69.6 ± 2.6	9
Baselines	**GraphMVP** [Liu *et al.*, 2022]	GNN	70.8 ± 0.5	79.3 ± 1.5	79.1 ± 2.8	76.0 ± 0.1	59.6 ± 3.9	73.0 ± 1.8	6
Baselines	**3D-infomax** [Stärk *et al.*, 2022]	GNN	69.1 ± 1.2	78.6 ± 1.9	62.7 ± 3.3	76.1 ± 1.3	60.7 ± 3.1	69.4 ± 2.2	11
Baselines	**MolCLR** [Wang *et al.*, 2022]	GNN	73.1 ± 1.6	81.5 ± 1.6	91.6 ± 2.7	77.3 ± 1.3	60.0 ± 2.9	76.7 ± 2.0	5
Baselines	**MoleBert** [Xia *et al.*, 2022]	GNN	71.9 ± 1.6	80.8 ± 1.4	78.9 ± 3.0	78.2 ± 0.8	51.4 ± 5.0	72.2 ± 2.4	8
Baselines	**Uni-Mol** [Zhou *et al.*, 2023]	Transformer	71.5 ± 1.4	**84.4 ± 2.1**	87.8 ± 2.6	78.3 ± 1.3	62.3 ± 5.6	76.9 ± 2.6	4
Baselines	**GROVER** [Rong *et al.*, 2020]	Transformer	65.1 ± 2.5	81.1 ± 2.3	74.0 ± 11.8	57.7 ± 4.3	56.8 ± 5.1	67.0 ± 5.2	12
Ours	M^2**LLM**($LLaMa$-3.1)	LLM	**77.0 ± 1.0**	77.8 ± 2.9	99.1 ± 0.4	77.2 ± 0.9	62.7 ± 0.4	78.8 ± 1.1	3
Ours	M^2**LLM**($Galactica$)	LLM	74.9 ± 0.74	80.0 ± 2.7	**99.4 ± 0.1**	**77.5 ± 0.7**	**62.8 ± 0.4**	**79.0 ± 0.9**	**2**
Ours	M^2**LLM**($OpenAI$)	LLM	**75.5 ± 1.3**	78.2 ± 0.9	**99.5 ± 0.1**	**79.5 ± 0.7**	**63.7 ± 0.3**	**79.3 ± 0.7**	**1**

Table 1: Results on molecular property classification tasks with scaffold split. Mean and standard deviation of ROC-AUC (%) from 10 random seeds are reported, with higher values indicate better performance. The top-2 performances on each dataset are shown in bold, with **bold** being the best result, and **bold** being the second best result.

Dataset can be found in Appendix A.3. In the case of the BACE dataset, M^2LLM achieves 80.0%, which, while competitive, remains below the highest baseline result of 85.0% achieved by the RF model. The potential reasons may be the BACE dataset assigns binary labels for molecular inhibitors of human β-secretase 1 (BACE-1), based on an arbitrary threshold of quantitative potency values (IC_{50}) set at 7 [Wu *et al.*, 2018]. However, potency values can vary significantly depending on the assay settings [Landrum and Riniker, 2024], lower potency values can still indicate strong inhibition of BACE-1 [Harding *et al.*, 2024]. We hypothesize that this arbitrary threshold and label ambiguity hinder LLMs' ability to reason effectively.

4.3 Performance on Regression Tasks

We evaluate M^2LLM on three regression tasks, as shown in Table 2. We report the RMSE for regression, where lower values signify better result. Results presented in Table 2 demonstrate the superior performance of our proposed framework compared to all baselines across three datasets. Specifically, M^2LLM demonstrates strong performance, achieving an RMSE of 2.01 on FreeSolv dataset, reducing the error by 15.5% compared to the best baseline value of 2.38. Furthermore, on ESOL dataset, it achieves an RMSE of 0.44, a remarkable 56.9% reduction in error compared to the best baseline value of 1.02. Additionally, on the Lipophilicity dataset, it achieves state-of-the-art results with an RMSE of 0.66, while our other variants demonstrate competitive performance against baseline models.

Model \ Dataset	ESOL(1) ↓	FreeSolv(1) ↓	Lipophilicity(1) ↓
RF + ECFP4	1.34 ± 0.01	4.36 ± 0.04	0.90 ± 0.00
AttrMask	1.11 ± 0.05	2.92 ± 0.03	0.73 ± 0.00
GraphCL	1.31 ± 0.07	3.60 ± 0.32	0.78 ± 0.02
GraphMVP	1.06 ± 0.02	2.95 ± 0.19	0.69 ± 0.01
3D-infomax	0.89 ± 0.04	2.83 ± 0.10	0.70 ± 0.02
MolCLR	1.31 ± 0.03	2.73 ± 0.08	0.74 ± 0.02
MoleBert	1.02 ± 0.03	3.08 ± 0.05	**0.68 ± 0.02**
Uni-Mol	1.55 ± 0.26	3.94 ± 0.50	1.19 ± 0.07
GROVER	1.13 ± 0.08	2.38 ± 0.40	0.91 ± 0.09
M^2**LLM**($LLaMa$-3.1)	**0.44 ± 0.01**	**2.01 ± 0.37**	0.73 ± 0.03
M^2**LLM**($Galactica$)	0.53 ± 0.25	2.39 ± 1.39	**0.66 ± 0.02**
M^2**LLM**($OpenAI$)	**0.48 ± 0.02**	**2.35 ± 0.63**	0.77 ± 0.01

Table 2: Results on Molecular Property Regression tasks with Scaffold Split. Mean and standard deviation of the Root Mean Square Error (RMSE) metric from 10 random seeds are reported, with lower scores indicating better performance. Average statistics of target labels are -3.46 for ESOL, -6.33 for FreeSolv, and 2.20 for Lipophilicity.

4.4 Multi-view Component Contribution Analysis

In this section, we analyze the contribution of each view component to the final decision, as illustrated in Figure 3, based on three classification datasets and three regression tasks. Interestingly, the molecular structure view component contributes more significantly to the classification tasks, whereas the molecular rule view component and molecular task view component play a larger role in the regression tasks. This suggests that classification tasks may benefit from a detailed representation of molecular structures, as these tasks often rely on recognizing specific structural features critical for distinguishing between categories. On the other hand, regression tasks, which predict continuous values such as molecular properties, appear to benefit more from the molecular rule and task views, which capture broader context and relationships. This demonstrates that our framework effectively automates the learning of appropriate weights for each component, dynamically optimizing the contribution of each view for individual molecules based on the task requirements.

From the LLM architecture perspective, we find that LLMs with the same architecture tend to exhibit similar component contributions across different datasets to the final predictions. However, on the ESOL and FreeSolv datasets, even though the LLMs heavily rely on one or two components, Galactica-6.7B and 30B demonstrate different behavior, despite having the same architectural design, these models exhibit different component contributions for their final decisions. This pat-

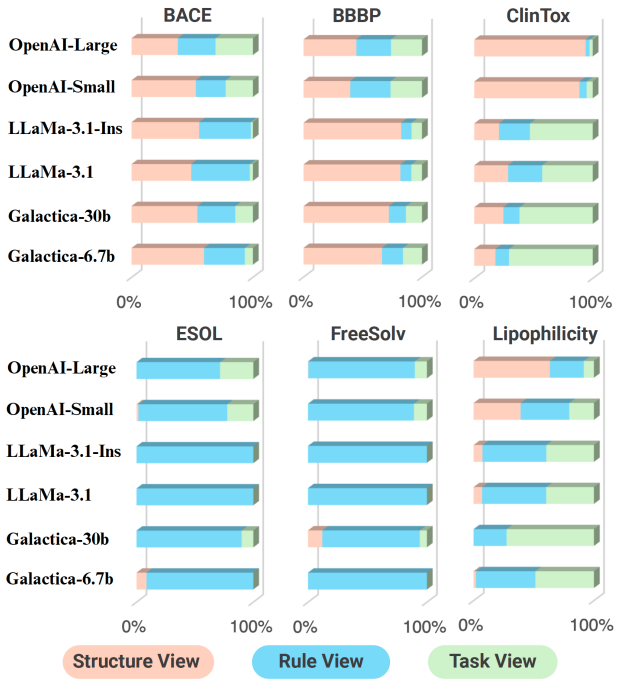

Figure 3: Multi-view Component Contribution Analysis. The contributions of the Molecular Structure View, Molecular Rules View, and Molecular Task View to the final score are calculated by averaging component weights for each SMILES representation across 10 random seeds.

tern highlights the flexibility and adaptability of our proposed multi-view representation learning framework.

Furthermore, the results for the ClinTox dataset, as reported in Table 1, demonstrate that we achieve near-perfect scores across all backbone settings. However, this component analysis reveals intriguing insights into how different model architectures rely on various components to make their predictions. The OpenAI models heavily depend on the molecular structure view, indicating that their decision-making process is primarily driven by structural understanding and extensive pre-trained knowledge. In contrast, the Galactica models rely more on task-specific components, likely because our task-specific thinking process is closely aligned with their pre-training dataset and methodology. The LLaMa-3.1 models demonstrate a relatively balanced utilization of all three components to make accurate predictions.

4.5 Effectiveness of Multi-view Representation

To better understand the performance gains afforded by our proposed multi-view representation, we first evaluate a baseline configuration using the SMILES-only representation. Specifically, only the SMILES string of a molecule is fed into the best-performing LLM model. This approach relies solely on the LLM's general understanding of a molecule and does not prompt the model to reason through the contextual diversity provided by a multi-view approach.

As shown in Figure 4, our proposed method consistently improved the scores across all six datasets. On the FreeSolv regression dataset, where a lower RMSE indicates better performance, we achieved a substantial reduction in prediction error, decreasing it from 4.29 to 2.01, representing a 53.3% improvement. This improvement underscores the effectiveness of our framework, particularly in molecular property regression tasks. Similarly, for other regression and classification tasks, such as ESOL and BBBP, our method demonstrated a measurable improvement compared to the SMILES-only baseline. This trend is also consistently observed for all other LLM backbone models as illustrated in Appendix A.4.

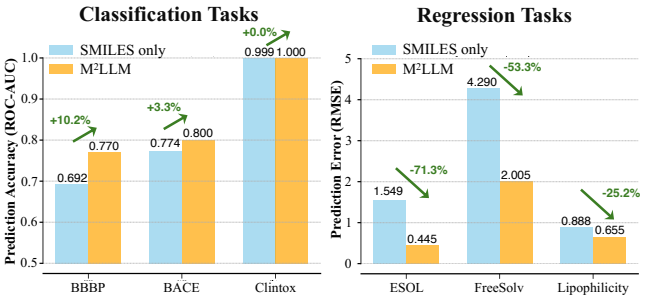

Figure 4: Comparison of the performance between M^2LLM and the SMILES-only representation across 10 random seeds on six datasets.

In the case of the ClinTox dataset, the SMILES-only baseline achieves nearly perfect results, with our framework offering a marginal improvement. This suggests that in tasks where the LLM already possesses sufficient understanding of the molecular domain through SMILES-based encoding alone, the additional views provide less pronounced benefits. This observation, nonetheless, further reinforces the strength of LLMs as text encoders for molecular property prediction.

Overall, these results highlight the key contribution of M^2LLM: the integration of complementary molecular views consistently enhances predictive performance compared to a single-view SMILES-only approach. By dynamically incorporating diverse representations, M^2LLM captures richer molecular features and demonstrates superior adaptability across tasks of varying complexity, firmly establishing itself as a state-of-the-art framework for molecular property prediction.

5 Conclusion

In this paper, we introduce M^2LLM, a multi-view learning framework that harnesses the capabilities of LLMs to generate rich molecular representations, enabling state-of-the-art performance in molecular property prediction. By utilizing the strong reasoning capabilities, extensive pre-trained knowledge, and powerful encoding abilities of LLMs, the framework delivers exceptional results across several benchmark tasks. Unlike methods that rely solely on SMILES as input, M^2LLM dynamically integrates multiple views to capture complex molecular features, enabling the learned representations to generalize effectively across diverse classification and regression tasks. These results underscore the transformative potential of M^2LLM in advancing molecular property prediction, offering a scalable and versatile solution for a wide range of applications in molecular science and beyond.

Acknowledgments

This research was partly funded by Australian Research Council (ARC) under grants FT210100097 and DP240101547.

References

[Breiman, 2001] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[Bu *et al.*, 2024] Weixin Bu, Xiaofeng Cao, Yizhen Zheng, and Shirui Pan. Improving augmentation consistency for graph contrastive learning. *Pattern Recognition*, 148:110182, 2024.

[Delaney, 2004] John S Delaney. Esol: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3):1000–1005, 2004.

[Drews, 2000] Jurgen Drews. Drug discovery: a historical perspective. *science*, 287(5460):1960–1964, 2000.

[Du *et al.*, 2024] Wenjie Du, Shuai Zhang, Jun Xia Di Wu, Ziyuan Zhao, Junfeng Fang, and Yang Wang. Mmgnn: A molecular merged graph neural network for explainable solvation free energy prediction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 5808–5816, 2024.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Fabian *et al.*, 2020] Benedek Fabian, Thomas Edlich, et al. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*, 2020.

[Gayvert *et al.*, 2016] Kaitlyn M Gayvert, Neel S Madhukar, and Olivier Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell chemical biology*, 23(10):1294–1301, 2016.

[Harding *et al.*, 2024] Simon D Harding, Jane F Armstrong, et al. The iuphar/bps guide to pharmacology in 2024. *Nucleic Acids Research*, 52(D1):D1438–D1449, 2024.

[Hu *et al.*, 2019] Weihua Hu, Bowen Liu, et al. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[Jeon and Kim, 2019] Woosung Jeon and Dongsup Kim. Fp2vec: a new molecular featurizer for learning molecular properties. *Bioinformatics*, 35(23):4979–4985, 2019.

[Kenton and Toutanova, 2019] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

[Koh *et al.*, 2024] Huan Yee Koh, Anh TN Nguyen, Shirui Pan, Lauren T May, and Geoffrey I Webb. Physicochemical graph neural network for learning protein–ligand interaction fingerprints from sequence data. *Nature Machine Intelligence*, pages 1–15, 2024.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, et al. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Kuhn *et al.*, 2016] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic acids research*, 44(D1):D1075–D1079, 2016.

[Landrum and Riniker, 2024] Gregory A Landrum and Sereina Riniker. Combining ic50 or k i values from different sources is a source of significant noise. *Journal of Chemical Information and Modeling*, 64(5):1560–1567, 2024.

[Liu *et al.*, 2022] Shengchao Liu, Hanchen Wang, et al. Pre-training molecular graph representation with 3d geometry. In *International Conference on Learning Representations*, 2022.

[Luo *et al.*, 2024] Yizhen Luo, Kai Yang, et al. Learning multi-view molecular representations with structured and unstructured knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2082–2093, 2024.

[Martins *et al.*, 2012] Ines Filipa Martins, Ana L Teixeira, et al. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

[Medsker *et al.*, 2001] Larry R Medsker, Lakhmi Jain, et al. Recurrent neural networks. *Design and Applications*, 5(64-67):2, 2001.

[Mirza *et al.*, 2024] Adrian Mirza, Nawaf Alampara, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024.

[Mobley and Guthrie, 2014] David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.

[OpenAI *et al.*, 2023] Josh OpenAI, Achiam, Steven Adler, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[OpenAI, 2024] OpenAI. New embedding models and api updates, 2024.

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

[Rollins *et al.*, 2024] Zachary A Rollins, Alan C Cheng, and Essam Metwally. Molprop: Molecular property prediction with multimodal language and graph fusion. *Journal of Cheminformatics*, 16(1):56, 2024.

[Rong *et al.*, 2020] Yu Rong, Yatao Bian, et al. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

[Ross *et al.*, 2022] Jerret Ross, Brian Belgodere, et al. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[Sadeghi *et al.*, 2024] Shaghayegh Sadeghi, Alan Bui, Ali Forooghi, Jianguo Lu, and Alioune Ngom. Comparative analysis of llama and chatgpt embeddings for molecule embedding. *arXiv preprint arXiv:2402.00024*, 2024.

[Shirasuna *et al.*, 2024] Victor Yukio Shirasuna, Eduardo Soares, et al. A multi-view mixture-of-experts based on language and graphs for molecular properties prediction. In *ICML 2024 AI for Science Workshop*, 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.

[Subramanian *et al.*, 2016] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of β-secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of chemical information and modeling*, 56(10):1936–1949, 2016.

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

[Wang *et al.*, 2019] Sheng Wang, Yuzhi Guo, et al. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436, 2019.

[Wang *et al.*, 2022] 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.*, 2024a] Luzhi Wang, Dongxiao He, He Zhang, Yixin Liu, Wenjie Wang, Shirui Pan, Di Jin, and Tat-Seng Chua. Goodat: towards test-time graph out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15537–15545, 2024.

[Wang *et al.*, 2024b] Luzhi Wang, Yizhen Zheng, Di Jin, Fuyi Li, Yongliang Qiao, and Shirui Pan. Contrastive graph similarity networks. *ACM Transactions on the Web*, 18(2):1–20, 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.

[Wu *et al.*, 2024] Man Wu, Xin Zheng, Qin Zhang, Xiao Shen, Xiong Luo, Xingquan Zhu, and Shirui Pan. Graph learning under distribution shifts: A comprehensive survey on domain adaptation, out-of-distribution, and continual learning. *arXiv preprint arXiv:2402.16374*, 2024.

[Xia *et al.*, 2022] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *The Eleventh International Conference on Learning Representations*, 2022.

[Yang *et al.*, 2019] Kevin Yang, Kyle Swanson, Wengong Jin, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

[Yu *et al.*, 2024] Jiajun Yu, Zhihao Wu, Jinyu Cai, Adele Lu Jia, and Jicong Fan. Kernel readout for graph neural networks. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 2505–2514, 2024.

[Yu *et al.*, 2025] Jiajun Yu, Yizhen Zheng, Huan Yee Koh, Shirui Pan, Tianyue Wang, and Haishuai Wang. Collaborative expert llms guided multi-objective molecular optimization. *arXiv preprint arXiv:2503.03503*, 2025.

[Zhang *et al.*, 2019] He Zhang, Hanlin Mo, You Hao, Qi Li, Shirui Li, and Hua Li. Fast and efficient calculations of structural invariants of chirality. *Pattern Recognit. Lett.*, 128:270–277, 2019.

[Zhang *et al.*, 2025] He Zhang, Bang Wu, Xiangwen Yang, Xingliang Yuan, Xiaoning Liu, and Xun Yi. Dynamic graph unlearning: A general and efficient post-processing method via gradient transformation. In *Proceedings of the ACM on Web Conference 2025*, pages 931–944, 2025.

[Zheng *et al.*, 2024a] Xin Zheng, Dongjin Song, Qingsong Wen, Bo Du, and Shirui Pan. Online gnn evaluation under test-time graph distribution shifts. In *The Twelfth International Conference on Learning Representations*, 2024.

[Zheng *et al.*, 2024b] Yan Zheng, Song Wu, Junyu Lin, Yazhou Ren, Jing He, Xiaorong Pu, and Lifang He. Cross-view contrastive fusion for enhanced molecular property prediction. *Proccedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

[Zheng *et al.*, 2024c] Yizhen Zheng, Huan Yee Koh, Maddie Yang, Li Li, Lauren T May, Geoffrey I Webb, Shirui Pan, and George Church. Large language models in drug discovery and development: From disease mechanisms to clinical trials. *arXiv preprint arXiv:2409.04481*, 2024.

[Zheng *et al.*, 2025] Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11, 2025.

[Zhou *et al.*, 2023] Gengmo Zhou, Zhifeng Gao, et al. Unimol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.