MolX: Enhancing Large Language Models for Molecular Understanding With A Multi-Modal Extension

Khiem Le[1], Zhichun Guo[1], Kaiwen Dong[1], Xiaobao Huang[1], Bozhao Nan[1], Roshni Iyer[2], Xiangliang Zhang[1], Olaf Wiest[1], Wei Wang[2], Ting Hua[1], Nitesh V. Chawla[1]
[1]University of Notre Dame, IN, USA
[2]University of California, Los Angeles, CA, USA

Abstract

Large Language Models (LLMs) with their strong task-handling capabilities have shown remarkable advancements across a spectrum of fields, moving beyond natural language understanding. However, their proficiency within the chemistry domain remains restricted, especially in solving molecule-related tasks. This challenge is attributed to their inherent limitations in comprehending molecules using only common textual representations, i.e. SMILES strings. In this study, we seek to enhance the ability of LLMs to comprehend molecules by equipping them with a multi-modal external module, termed MolX. Instead of directly using SMILES strings to represent a molecule, we utilize specific encoders to extract fine-grained features from both SMILES string and 2D molecular graph representations for feeding into an LLM. A hand-crafted molecular fingerprint is incorporated to leverage its embedded domain knowledge. To establish an alignment between MolX and the LLM's textual input space, the model in which the LLM is frozen, is pre-trained with a strategy including a diverse set of tasks. Experimental evaluations show that our proposed method outperforms baselines across 4 downstream molecule-related tasks ranging from molecule-to-text translation to retrosynthesis, with and without fine-tuning the LLM, while only introducing a small number of trainable parameters—0.53% and 0.82%, respectively.

CCS Concepts

• **Computing methodologies** → **Learning paradigms**; • **Applied computing** → **Bioinformatics**.

Keywords

Large Language Models, Multi-Modal Learning, Molecular Understanding, Molecule-Related Tasks

ACM Reference Format:
Khiem Le[1], Zhichun Guo[1], Kaiwen Dong[1], Xiaobao Huang[1], Bozhao Nan[1], Roshni Iyer[2],, Xiangliang Zhang[1], Olaf Wiest[1], Wei Wang[2], Ting Hua[1], Nitesh V. Chawla[1]. 2025. MolX: Enhancing Large Language Models for Molecular Understanding With A Multi-Modal Extension. In *Proceedings of 2025 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (MLoG-GenAI@KDD '25).* ACM, Toronto, ON, Canada, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

MLoG-GenAI@KDD '25,

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large Language Models (LLMs) have demonstrated impressive performances across a wide array of fields. Extending beyond the boundaries of natural language understanding, LLMs have facilitated various scientific disciplines [38, 39]. LLMs have recently been investigated for augmenting research in chemistry as an alternative approach to the traditional supervised learning approach [1, 6].

Despite their strong task-handling capabilities, LLMs still struggle with the chemistry domain, as evidenced by their limited performances on various professional molecule-related tasks [13, 51]. Llama-2 [40], performs unsatisfactorily on the molecule-to-text translation tasks such as molecule description generation and IUPAC name generation, providing the correct answer only half as often as supervised learning models. Additionally, such LLM fails to predict molecular properties even using expert-designed prompts. One potential cause of this challenge has been figured out that most existing LLMs represent molecules only by their common textual representations, i.e., SMILES strings [45], and process them in a paradigm similar to texts [13, 22], as illustrated in Figure 1a. While convenient, several issues make it challenging for LLMs to comprehend molecules solely by interpreting SMILES strings. Firstly, LLMs lack an inherent understanding of SMILES strings and blindly treat them as sequences of separate characters relying on their byte-pair encoding tokenizers [36], which break SMILES strings into smaller pieces in ways that do not represent the chemical principles behind these strings. Without an understanding of these principles, it is difficult for LLMs to capture molecular structure from SMILES strings due to inaccuracies such as incorrect transcription of complex aromatic systems or the absence of hydrogens and other atoms [41], as shown in Figure 1b and Figure 1c.

There have been some early attempts to enhance LLMs for solving molecule-related tasks. Su et al. [37] employed a GNN-based graph encoder to extract features from the molecule's 2D molecular graph and directly input such features into the LLM to perform molecule-to-text translation tasks. Developed from that idea, Li et al. [22] input features extracted from the 2D or 3D molecular graph into the LLM through an intermediate projector, which is previously aligned with the LLM's textual input space by a pre-training stage. Although bridging the gap between the 2D or 3D molecular graph and the LLMs, previous approaches are ineffective in using the information contained in a SMILES string, as well as handcrafted molecular descriptors, which have advantages over 2D or 3D molecular graph [8, 19]. This might lead to suboptimal performances. Existing methods are only optimized for a limited

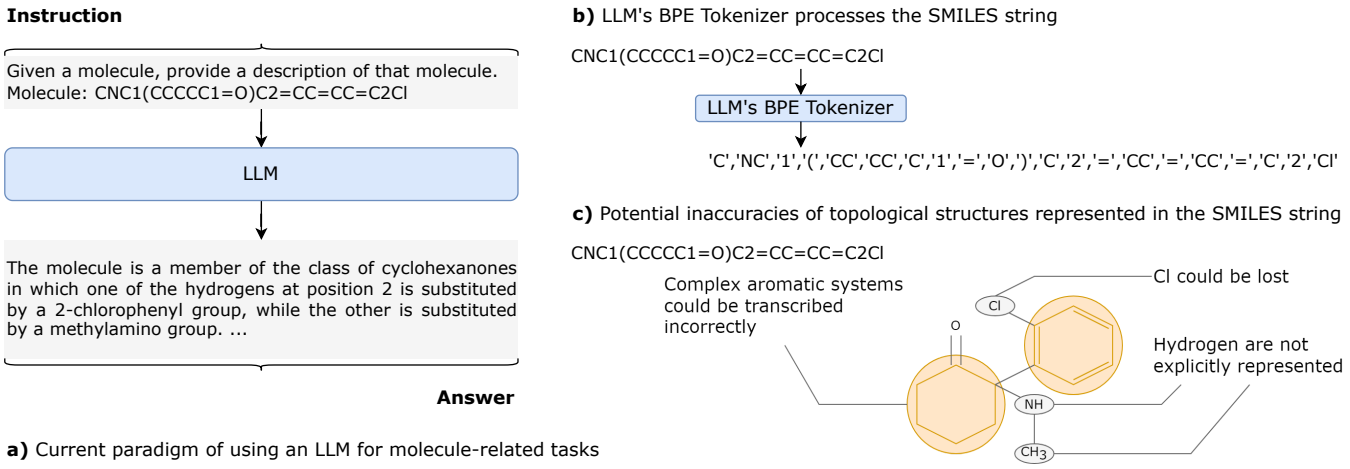

Figure 1: Current paradigm of using an LLM for molecule-related tasks and its issues.

number of chemistry-related tasks, omitting other crucial tasks such as molecular property prediction, molecular optimization, or retrosynthesis.

In this study, we introduce a novel framework for enhancing LLMs to capture molecules from multiple representations, thus improving their performances on various molecule-related tasks. Our proposed framework consists of two main components which are a multi-modal external module, namely MolX, equipped with the LLMs, and a versatile pre-training strategy for aligning MolX into the LLMs' textual input space. We first utilize a pre-trained BERT-like [9] SMILES encoder to extract features from the SMILES string instead of directly using it to represent the molecule. Because of its initial pre-training, the SMILES encoder works with its tokenizer to capture long-range dependencies encoded in the SMILES string. We simultaneously utilize a pre-trained GNN-based graph encoder to extract features from the molecule's 2D molecular graph, capturing its topological structures. In addition to features extracted from raw representations, i.e., SMILES string and 2D molecular graph, a handcrafted molecular fingerprint [26] containing domain knowledge is incorporated in a weighted scheme of MolX. Finally, the modelin which the LLM is frozen undergoes pre-training strategy with a diverse set of tasks, providing the model with information about the molecules. This process provides an alignment between MolX and the LLM's textual input space. Figure 2 shows an overview of our proposed method.

Our experimental results demonstrate that the proposed method outperforms baselines by a statistically significant margin on various downstream molecule-related tasks in two different model configurations, with and without fine-tuning the LLM. It is worth noting that MolX can act as a plug-in module to the LLMs for enhancing the performances on molecule-related tasks while fully preserving its general-purpose usage in other domains.

To summarize, our contributions are outlined as follows:

- We introduce a novel framework enhancing LLMs to comprehend molecules, thus improving their performances on various molecule-related tasks. The LLMs are equipped with a multi-modal external module, MolX, to extract features from both SMILES string and 2D molecular graph representations, as well as leverage a handcrafted molecular fingerprint.
- A pre-training strategy including a diverse set of tasks, is applied to establish an alignment between MolX and the LLMs' textual input space. This process advances the models' ability of molecular understanding, as well as instruction following.
- Extensive experimental evaluations demonstrate that our proposed method outperforms baselines by a substantial margin on a diverse range of downstream molecule-related tasks in two different model configurations, with and without fine-tuning the LLM.

2 Related Work

In this section, we provide a review of the literature related to molecular learning via language modeling and leveraging LLMs for solving molecule-related tasks.

2.1 Molecular Learning

Molecules form the basis of chemistry and molecular learning has been a long-standing problem in cheminformatics [3, 29, 30, 48]. Traditionally, molecular fingerprints such as Morgan fingerprint [26] or ECFP [32] serve as one of the most important descriptors for molecules, encoding a molecule into a fixed bit string, where each bit indicates the presence of a certain substructure. With the rapid development of language modeling, textual representations such as SMILES strings have become widely used [45]. Studying molecule property prediction tasks, Wang et al. [43] introduced SMILES-BERT, a BERT-like model [9] that is pre-trained with the masked language modeling mechanism (MLM) on a large-scale set of unlabeled molecules. Wang et al. [42] proposed using chemical

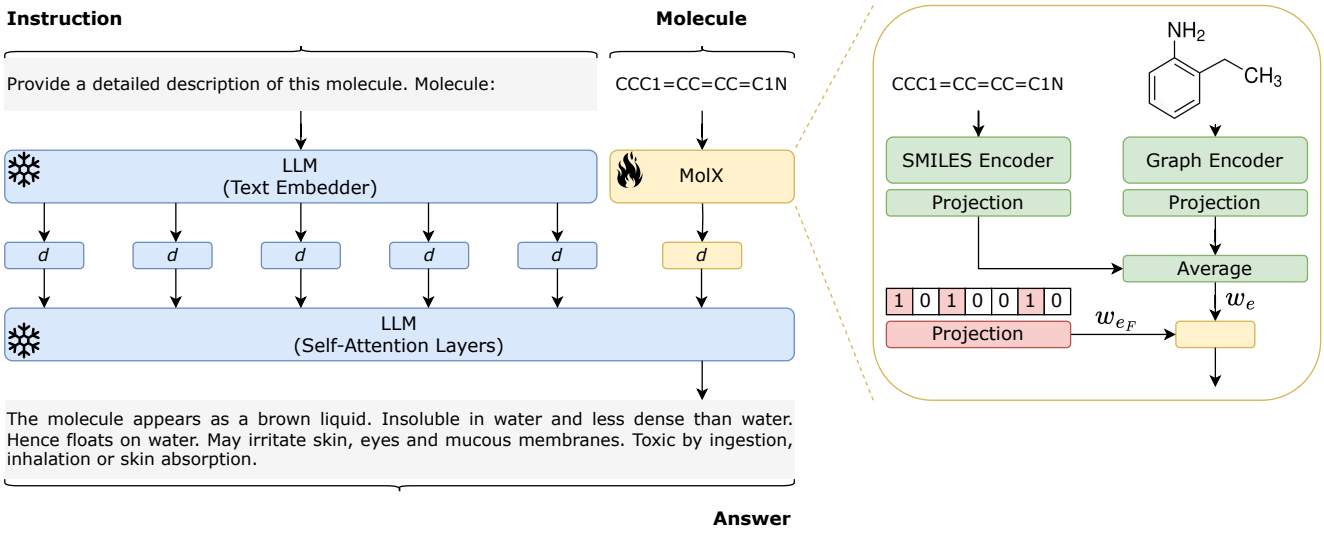

Figure 2: An overview of our proposed method with the main pre-training task.

reactions to assist the pre-training. Ahmad et al. [2] proposed using auxiliary tasks with more domain relevance for chemistry such as predicting computed properties of molecules, supporting MLM. Irwin et al. [18] investigated the challenging sequence-to-sequence tasks such as retrosynthesis and introduced Chemformer. Zhong et al. [53] proposed the root-aligned SMILES (R-SMILES), adopting a tighter representation for those tasks. Edwards et al. [11] studied molecule-to-text translation tasks and vice versa and proposed MolT5, which is pre-trained with the multi-lingual MLM, considering SMILES strings as a conventional language. Lu and Zhang [25] and Christofidellis et al. [7] presented ChemT5 and Text+ChemT5, unifying all sequence-to-sequence tasks. Several studies [14, 23] demonstrated that fusing the molecule's 2D molecular graphs with language modeling provides complementary benefits to molecular learning, improving performance on tasks such as molecule property prediction. With rising use across a wide array of fields, including chemistry [1, 6], LLMs have emerged as an evolution of the traditional language modeling approach for molecular learning.

2.2 LLMs for Molecule-Related Tasks

Several studies have evaluated LLM applications in chemistry. Castro Nascimento and Pimentel [6] explored how well ChatGPT "understands" chemistry by posing five student-level tasks in different subareas of chemistry and noticed moderate performance. Zhao et al. [51] investigated the molecule property prediction task and showed that LLMs relied on memorized information rather than true understanding for making predictions, which limits their applications to new types of molecules required in practical applications. Guo et al. [13] benchmarked several published LLMs on eight molecule-related tasks. Empirical results reveal that LLMs such as Llama-2 [40] that were widely used at the time typically fail to perform challenging tasks of molecule-to-text translation or predict molecule activity for high-level properties even using expert-designed prompts. A potential reason behind this challenge has been identified that most existing LLMs represent molecules only by their common textual representations, i.e., SMILES strings, which LLMs have a limited understanding of. In response to these findings, Su et al. [37] propose MoMu to enhance LLMs by applying a GNN-based graph encoder to extract features from the molecule's 2D molecular graph and input such features into the LLM for performing molecule-to-text translation tasks. Li et al. [22] proposed 2D and 3D MoLM to leverage an intermediate projector for feeding features extracted from the 2D or 3D molecular graph into the LLM, which is previously aligned with the LLM's textual input space by a pre-training stage. Despite improvements by bridging the gap between the 2D or 3D molecular graph and the LLMs, the importance of representation other than SMILES strings such as handcrafted molecular descriptors are underexplored. Existing methods are only optimized for a limited set of molecule-related tasks, how well the enhanced LLMs perform on other tasks such as molecular property prediction, molecule optimization, or retrosynthesis is not well understood.

3 Methodology

We propose a framework enhancing LLMs to comprehend molecules from multiple representations, consisting of two main components, a multi-modal external module and a novel pre-training strategy. Here, we present the details of these components.

3.1 Model Architecture

The proposed MolX, which is equipped with a base LLM, consists of two key designs: 1) Trainable encoders, focusing on encoding raw representations of a molecule, i.e., SMILES string and 2D molecular graph; 2) A weighted scheme to incorporate a handcrafted molecular fingerprint.

Trainable Encoders. We define a molecule as m and consider m_S and m_G to depict its SMILES string and 2D molecular graph, respectively. While m_S is simply a sequence of ASCII characters, m_G is considered as $m_G = \{\mathcal{V}, \mathcal{E}\}$, where each node in $\mathcal{V}$ indicates an atom and each edge in $\mathcal{E}$ indicates a chemical bond. Also, $\boldsymbol{X} \in$

$\mathbb{R}^{|\mathcal{V}|\times N}$ is the attribute matrix of m_G where $x_n = \boldsymbol{X}[n,:]^T$ is the N-dimensional attribute vector of the node $v_n \in \mathcal{V}$.

To encode the SMILES string m_S, we adopt a pre-trained BERT-like [9] SMILES encoder, ChemBERTa [2], which is constructed by stacking multiple Transformer layers. ChemBERTa, denoted as E_S, is pre-trained on a large-scale set of unlabeled molecules with MLM, enabling it to capture long-range dependencies identified in the SMILES string. An average is taken over outputs of E_S to obtain an embedding vector for m_S, which is then projected to the hidden dimension d of the base LLM by a multi-layer perceptron f_S:

$$e_S = f_S(\text{Average}(\{t_i, t_i \in E_S(m_S)\})) \in \mathbb{R}^d. \tag{1}$$

To encode the 2D molecular graph m_G, we adopt a pre-trained GNN-based graph encoder, ChemGraphCL [49], which is constructed based on an emerging message-passing GNN, GIN [16]. ChemGraphCL, denoted as E_G, is pre-trained on a large-scale set of unlabeled molecules with a contrastive learning strategy [31] and thus able to capture the topological structures of the molecule from its 2D molecular graph. Starting from the initial x_n, after multiple layers of message propagation, E_G produces an updated attribute vector h_n for the node $v_n \in \mathcal{V}$. Then, an average is taken over all node-level attribute vectors to obtain an embedding vector for m_G, which is projected to the hidden dimension d of the base LLM by a multi-layer perceptron f_G:

$$e_G = f_G(\text{Average}(\{h_n, h_n \in E_G(m_G)\})) \in \mathbb{R}^d. \tag{2}$$

e_S and e_G are then averaged to establish a unified embedding vector $e \in \mathbb{R}^d$.

Molecular Fingerprint Incorporation. Molecular fingerprints are some of the most important descriptors of molecules due to the encoded domain knowledge. While SMILES strings and 2D molecular graphs capture global information about the molecule, molecular fingerprints capture information about the local atomic environments and neighborhoods, explicitly encoding the presence of specific substructures [10]. Unfortunately, molecular fingerprints are not often used in deep learning models even though they have been shown to be valuable for specific tasks such as molecule property prediction [47]. We seek to exploit their benefits by incorporating the popular Morgan fingerprint [26] into the unified embedding vector e from trainable encoders described above. RDKit [20] is utilized to compute the Morgan fingerprints with a radius of 2 from the molecule m, which is then projected to the hidden dimension d of the base LLM by a multi-layer perceptron f_F. The incorporation scheme works as follows:

$$\begin{aligned} e &= w_e \cdot e + w_{e_F} \cdot e_F, \\ &\text{where } e_F = f_F(\texttt{MorganFP}(m)), \end{aligned} \tag{3}$$

where w_e and w_{e_F} are trainable parameters introduced for providing the model sufficient flexibility to incorporate the Morgan fingerprint into e.

3.2 Pre-training Strategy

There is a noticeable misalignment in the latent spaces of MolX and the base LLM where the former encodes molecules while the latter has a textual input space. Therefore a cross-space alignment stage is needed. This is accomplished by feeding the embedding vector from MolX into the LLM as a soft token. We propose to pre-train

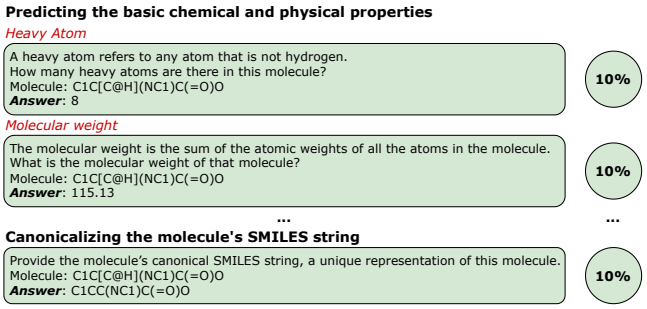

Figure 3: Examples of auxiliary tasks in our instruction-based pre-training strategy.

the MolX-enhanced LLM with a diverse set of tasks including a molecule-to-text translation task, i.e., molecule description generation, accompanied by several auxiliary tasks. It is worth noting that while MolX is trainable, the base LLM is kept frozen during pre-training. This setting maintains the LLM's inherent generalizability, forcing MolX to produce embedding vectors that are suited in the LLM's textual input space and can be effectively understood by the LLM to generate accurate answers. This allows the LLM to function normally on general domains by using MolX as a plug-in module for the handling of molecule-related tasks.

Multi-Task Dataset. To conduct the pre-training, we utilize the pre-train subset of PubChem [22], a dataset that contains 300k molecule-description pairs [1] for the molecule description generation task. By using this task as an objective, MolX is encouraged to produce meaningful embedding vectors, so that the LLM can caption molecules with their substructures and properties accurately, as illustrated in Figure 2. Although this dataset collected from a reliable source, descriptions in the dataset retain several limitations that might hinder the model's ability of molecular understanding. The average number of words in the dataset's descriptions is roughly 20, which is insufficient to describe a molecule. Additionally, some of the dataset's descriptions are noisy and uninformative [22]. Therefore, to assist the molecule description task, we design a set of auxiliary tasks including predicting the basic chemical and physical properties of molecules such as the number of heavy atoms or molecular weight. We select a set of 10 low-level properties that are available for easy collection from PubChem and present comprehensive information about the molecules. Further, leveraging the fact that a molecule can be represented by multiple valid SMILES strings [4], we utilize one more special auxiliary task which is canonicalizing the molecule's SMILES string. This task enhances the model's understanding of chemical laws behind SMILES strings. To keep the pre-training stage controllable, 10% of the dataset is used for each auxiliary task. Examples of proposed auxiliary tasks are shown in Figure 3 and details are in Appendix A.

Instruction-based Pre-training. LLMs tend to exhibit hallucinations in the domain of chemistry [13], generating unexpected answers regarding a molecule. Hence, we enrich our pre-training dataset by designing an informative instruction for each task. We then employ instruction-based pre-training [28, 34], enhancing the

[1] https://pubchem.ncbi.nlm.nih.gov

Table 1: Experimental results for molecule-to-text translation.

	Model	Description Generation						IUPAC Name Generation					
		BLE-2↑	BLE-4↑	ROG-1↑	ROG-2↑	ROG-L↑	MET↑	BLE-2↑	BLE-4↑	ROG-1↑	ROG-2↑	ROG-L↑	MET↑
Infer-only	Llama-2-7B	03.64	02.98	18.28	04.26	12.87	16.21	05.55	01.81	05.40	00.23	04.39	10.30
	Llama-2-7B + **MolX**	**08.22**	**06.40**	**30.82**	**21.69**	**28.94**	**21.77**	**10.67**	**04.76**	**14.61**	**01.24**	**11.47**	**18.54**
LoRA FT	Llama-2-7B	27.54	21.24	36.50	21.33	28.99	31.69	51.43	36.94	48.54	20.57	40.53	53.38
	Llama-2-7B + MoMu	27.68	21.50	36.76	21.42	29.23	31.86	51.70	37.38	48.89	20.65	40.87	53.66
	Llama-2-7B + MoLM-2D	27.95	21.77	38.66	22.99	30.92	33.69	52.32	37.65	51.77	21.83	43.62	57.10
	Llama-2-7B + MoLM-3D	29.82	22.39	39.12	23.62	32.64	34.34	55.70	38.93	52.03	22.78	45.63	57.84
	Llama-2-7B + **MolX**	**31.40**	**24.25**	**44.20**	**28.96**	**38.76**	**39.55**	**56.88**	**45.01**	**55.45**	**30.14**	**48.19**	**59.35**
	LlaSMol-7B	26.71	18.06	38.75	22.77	33.32	32.63	49.48	36.33	52.38	28.53	45.20	58.48
	ChemDFM-13B	13.02	08.30	20.42	11.31	17.93	18.44	39.33	22.83	37.61	09.49	28.68	45.99
Full FT	MolT5-Large	25.87	17.28	34.07	16.42	23.41	28.04	50.88	38.69	45.89	21.11	33.03	44.82
	MolT5-Large + MoMu	26.34	18.01	34.75	16.86	24.76	28.73	51.81	40.32	46.81	21.68	34.93	45.92

model's ability of instruction following. Formally, we first define $p(.)$ as the textual distribution parameterized by the base LLM. The base LLM is decomposed into two subparts, the text embedder F_{emb} and self-attention layers F_{att}, in which the text embedder F_{emb} converts an instruction of a task into a list of T tokens $Z = [z_1, z_2, .., z_T]$. Given a molecule m and its label y for the given task, after the embedding vector e is extracted from MolX, the auto-regressive loss for pre-training is defined as:

$$\begin{aligned}\mathcal{L}_{reg} &= -\log p(y|F_{att}(z_1, z_2, .., z_T, e)) \\ &= -\sum_{l=1}^{L} \log p(y_l|F_{att}(z_1, z_2, .., z_T, e), y_1, ..., y_{l-1}),\end{aligned} \tag{4}$$

where L is the length of the label y for the given task.

4 Experiments

In this section, we conduct experiments on various downstream molecule-related tasks including molecule-to-text translation, molecule property prediction, molecule optimization, and retrosynthesis, to demonstrate the effectiveness of our proposed method. Throughout experiments, we utilize Llama-2 [40] with 7B parameters as our base LLM to leverage its text generation capability and internal chemistry knowledge. We consider two different model configurations for the evaluation: I) Inference-only: The model is frozen after pre-training for direct question answering on downstream tasks, evaluating the model's generalizability without fine-tuning; II) LoRA fine-tuning: The model is fine-tuned on downstream tasks using a parameter-efficient technique, LoRA [17], verifying the model's adaptability in scenarios where downstream data are available. In addition to direct comparison with previous related works including MoMu [37], as well as 2D and 3D MoLM [22], we also compare with competitive supervised learning models in each task. For further reference, we evaluate two recently introduced generalist chemical LLMs derived from Llama-2 [40] that are tailored for molecule-related tasks, i.e., LlaSMol-7B [50] and ChemDFM-13B [52]. The experimental settings and hyper-parameters are provided in Appendix B.

4.1 Molecule-to-Text Translation

We first consider the molecule-to-text translation tasks, i.e., molecule description generation and IUPAC name generation. These tasks reflect the general molecular understanding of the model and have crucial applications, enabling humans to gain an overview of a molecule. We conduct experiments on the downstream subset of the PubChem dataset [22], which has 15k high-quality molecule-description pairs and is separate from the pre-train one. We opt not to use the CheBI-20 dataset [11] because it is also sourced from PubChem and can be viewed as an older version of the used dataset. Following [11, 22], we adopt BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR as evaluation metrics.

Table 1 presents experimental results for these tasks across 6 different metrics. Based on the Inference-only results, we observe that the proposed framework significantly enhances the base LLM for direct question answering on both tasks without fine-tuning. In the scenario of LoRA fine-tuning, the MolX-enhanced LLM demonstrates superior performance compared to baselines with the highest scores on all metrics, especially for ROUGE-based and METEOR metrics which might be attributed to the proposed versatile pre-training strategy that provides the model with comprehensive information about the molecules. The approach of fine-tuning the LLM to establish multi-modal models shows better performances than generalist chemical LLMs, i.e., LlaSMol-7B [50] and ChemDFM-13B [52], as well as competitive supervised learning models such as MolT5 [11] and its MoMu-enhanced one [37].

4.2 Molecule Property Prediction

Besides the overall understanding, we assess the model's perception of molecular properties by conducting experiments on the molecule property prediction task. This task involves approximating quantitative attributes such as solubility or determining the activity for assays of a molecule. We employ the MoleculeNet dataset [46] with 8 different subsets including ESOL, FreeSolv, Lipophilicity, MUV, HIV, BACE, BBBP, and Tox21. As evaluation metrics, RMSE is used for regression subsets, and Accuracy and F1 are used for classification, following [50]. Figure 4 shows an example of this task.

Experimental results in Table 2 show that MolX improves performances of the base LLM in both model configurations. Especially for Inference-only results, MolX remarkably narrows approximation errors. Additionally, MolX enhances the model's ability of instruction following, generating expected answers without LLMs's favorite phrases. In addition to LoRA fine-tuned models, we consider ChemGraphCL [49] which serves as the GNN-based graph encoder in MolX, ensuring an adequate comparison. We observe that the MolX-enhanced LLM achieves the best scores in 6 out of 8 subsets of the MoleculeNet dataset and is the second-best in the other 2. Notably, properties in the MoleculeNet dataset are unseen

Table 2: Experimental results for molecule property prediction.

	Model	ESOL RMSE↓	FreeSolv RMSE↓	Lipophilicity RMSE↓	MUV ACC↑ | F1↑	HIV ACC↑ | F1↑	BACE ACC↑ | F1↑	BBBP ACC↑ | F1↑	Tox21 ACC↑ | F1↑
Infer-only	Llama-2-7B	58.719	357.371	222.426	0.110 | 0.100	0.135 | 0.129	0.522 | 0.362	0.485 | 0.351	0.090 | 0.084
	Llama-2-7B + **MolX**	**4.929**	**9.692**	**1.605**	**0.827** | **0.454**	**0.807** | **0.484**	**0.530** | **0.524**	**0.588** | **0.516**	**0.622** | **0.459**
LoRA FT	Llama-2-7B	2.061	4.203	0.956	0.984 | 0.572	0.960 | 0.610	0.612 | 0.584	0.603 | 0.564	0.740 | 0.578
	Llama-2-7B + MoMu	2.112	4.214	0.998	0.992 | 0.576	0.968 | 0.614	0.618 | 0.587	0.612 | 0.574	0.746 | 0.582
	Llama-2-7B + MoLM-2D	1.521	3.161	0.898	0.992 | 0.588	0.968 | 0.627	0.631 | 0.599	0.624 | 0.586	0.746 | 0.594
	Llama-2-7B + MoLM-3D	1.095	2.119	0.780	0.992 | 0.600	0.968 | 0.640	0.644 | 0.587	0.637 | 0.574	0.746 | 0.606
	Llama-2-7B + **MolX**	**0.967**	2.371	0.808	**0.994** | **0.609**	**0.972** | **0.649**	**0.704** | **0.697**	**0.666** | **0.650**	**0.748** | **0.616**
	LlaSMol-7B	1.871	6.047	1.361	0.829 | 0.434	0.968 | 0.492	0.467 | 0.318	0.529 | 0.346	0.608 | 0.475
	ChemDFM-13B	8.476	9.686	2.180	0.923 | 0.483	0.952 | 0.534	0.564 | 0.518	0.522 | 0.505	0.677 | 0.529
Full FT	ChemGraphCL	1.231	2.951	0.822	0.992 | 0.589	0.968 | 0.628	0.659 | 0.657	0.638 | 0.629	0.746 | 0.596
	ChemGraphMVP	1.091	**2.106**	**0.718**	0.993 | 0.590	0.971 | 0.630	0.691 | 0.689	0.647 | 0.638	0.747 | 0.597

Instruction

Solubility (logS) can be approximated by negative LogP -0.01 * (MPt – 25) + 0.5.
What is the logS of this molecule?
Molecule: Cc1cc(=O)[nH]c(=S)[nH]1
Please answer the question with a numerical value only.

Answer

Llama-2-7B : Sure, based on the provided SMILES string, estimated logS of the molecule is **-0.49**
Llama-2-7B + **MolX** : **-2.3479**

Ground Truth **: -2.44**

Figure 4: An example of molecule property prediction.

from the pre-training stage, showing the strong adaptability of our proposed method on unseen downstream tasks.

4.3 Molecule Optimization

Molecule optimization [15] is a more challenging task to assess the model's perception of molecular properties and the understanding of chemical laws behind SMILES strings. This task aims to modify a molecule toward a target property profile and the model is expected to generate the SMILES string of the modified molecule. The used dataset, ChEMBL-02 [15], contains 200k molecule pairs from ChEMBL database [12] with changes in properties, i.e., solubility, clearance, and LogD. Following [11], we adopt BLEU-2, Levenshtein, Morgan fingerprint-based Similarity, and Validity as evaluation metrics. Figure 5 shows an example of this task.

Experimental results for this task are shown in Table 3. For inference-only results, MolX not only boosts the performances of the base LLM to an acceptable level but also reduces hallucinations with chemically unreasonable SMILES strings, which are typically found when LLMs generate SMILES strings [13]. As an example in Figure 5, although still imperfect, the MolX-enhanced LLM recognized that the fluorine atom is the key modification. In the LoRA fine-tuning scenario, the MolX-enhanced LLM outperforms baselines including robust supervised learning models, Chemformer [18] and ReactionT5 [33] in most metrics.

4.4 Retrosynthesis

Retrosynthesis is a crucial task in chemistry [27]. This task involves a reverse extrapolation from a molecule to possible reactants used in its synthesis. The model is expected to generate SMILES strings of reactants separated by a '.'. We use the USPTO-50k dataset [35], containing 50k reactions for conducting experiments. Following [11], we adopt evaluation metrics similar to those used for the molecule optimization task. Figure 6 shows an example of this task.

Table 3: Experimental results for molecule optimization.

	Model	BLE-2↑	Leven↓	FTS↑	Valid↑
Infer-only	Llama-2-7B	08.49	666.70	-	00.00
	Llama-2-7B + **MolX**	**30.87**	**88.66**	**0.3732**	**07.27**
LoRA FT	Llama-2-7B	72.32	17.34	0.5715	91.31
	Llama-2-7B + MoMu	63.78	22.20	0.4659	92.59
	Llama-2-7B + MoLM-2D	73.16	17.32	0.6010	93.20
	Llama-2-7B + MoLM-3D	73.83	16.99	0.5834	94.05
	Llama-2-7B + **MolX**	**74.32**	**16.82**	**0.6113**	94.29
	LlaSMol-7B	34.95	39.50	0.5431	**99.85**
	ChemDFM-13B	32.94	50.65	0.5302	43.08
Full FT	Chemformer	66.60	20.86	0.5690	99.36
	ReactionT5-Large	73.45	18.91	0.6058	99.81

From experiential results presented in Table 4, we can observe that MolX improves the Inference-only results of the base LLM and alleviates the hallucinations with a similar effect as the molecule optimization task. As an example in Figure 6, the MolX-enhanced LLM recognized the first reactant and slightly erred the second one with the lack of an isocyanate group O=C=N. In the LoRA fine-tuning scenario, the MolX-enhanced LLM surpasses baselines and robust supervised learning models, Chemformer [18] and ReactionT5 [33] in most metrics.

5 Discussion

Here we discuss the limitations of our work and future directions. Firstly, we are aligning MolX into the LLM via a soft token, which is simple but effective. Although we are aware of advanced cross-space alignment techniques such as Q-Former [21], we opt not to employ them since they require a large number of high-quality molecule-description pairs and an extra pre-training stage, leading to high computational costs. A better alignment technique tailored for molecule-related tasks needs to be explored. Moreover, throughout experiments, we show the limitations of current generalist chemical LLMs, therefore, a novel generalist chemical LLM enhanced with MolX should be developed. LLMs also have been demonstrated to have intriguing abilities like In-context Learning [5] or Chain-of-Thought [44]. Leveraging these abilities for molecule-related tasks is a potential direction.

6 Conclusion

In this paper, we propose a novel framework enhancing LLMs to comprehend molecules, thus, improving their performances on molecule-related tasks. The LLMs are equipped with a multi-modal external module, MolX, which is aligned with their textual input

Instruction

Modify the molecule to create a new one such that
the solubility is unchanged, the clearance is unchanged, and a change in LogD within the interval (0.1, 0.3].
Molecule: Fc1ccc(C2(c3nnc4n3CCCCCC4)CCCC2)cc1

Answer

Llama-2-7B : Fc1ccc(C2(c3nnc4n3CCCCCC4)CCCC2)cc1-O-[CH(CH3)CH2OH]-
Llama-2-7B + **MolX** : CCC1ccc(C2(C3nnnC4n3CCCCCC4)CCCC2)cc1

Ground Truth **:** Cc1ccc(C2(c3nnc4n3CCCCCC4)CCCC2)cc1

Figure 5: An example of molecule optimization.

Table 4: Experimental results for retrosynthesis.

	Model	BLE-2↑	Leven↓	FTS↑	Valid↑
Infer-only	Llama-2-7B	10.10	468.74	-	00.00
	Llama-2-7B + **MolX**	**36.73**	**62.33**	**0.4041**	**13.71**
LoRA FT	Llama-2-7B	80.37	16.22	0.6981	89.27
	Llama-2-7B + MoMu	70.88	20.77	0.5691	90.53
	Llama-2-7B + MoLM-2D	82.05	15.90	0.7126	91.13
	Llama-2-7B + MoLM-3D	81.31	16.21	0.7341	90.31
	Llama-2-7B + **MolX**	**82.59**	**15.74**	**0.7466**	92.19
	LlaSMol-7B	50.09	31.28	0.7351	**99.65**
	ChemDFM-13B	39.93	57.48	0.5380	14.04
Full FT	Chemformer	74.01	19.51	0.6951	97.14
	ReactionT5-Large	81.63	17.69	0.7400	97.58

Instruction

Provide SMILES strings of possible reactants used in the molecule's synthesis.
The reactants should be split by '.'.
Molecule: O=C(NCCCl)Nc1cccc(Br)n1

Answer

Llama-2-7B : C6H5Br + NH3 + NaCN + HCl + NaN3 + H2O
Llama-2-7B + **MolX** : Nc1cccc(Br)n1.O=C(NCCCl)N

Ground Truth **:** Nc1cccc(Br)n1.O=C=NCCCl

Figure 6: An example of retrosynthesis.

Table 5: Numbers of trainable parameters in experiments.

	Model	# Trainable Params	
		Pre-training↓	Downstream↓
LoRA FT	Llama-2-7B	0.0M (0.00%)	20.5M (0.30%)
	Llama-2-7B + MoMu	2.0M (0.00%)	22.5M (0.30%)
	Llama-2-7B + MoLM-2D	120.0M (1.74%)	120.0M (1.74%)
	Llama-2-7B + MoLM-3D	120.0M (1.74%)	120.0M (1.74%)
	Llama-2-7B + **MolX**	36.1M (0.53%)	56.6M (0.82%)
	LlaSMol-7B	0.0M (0.00%)	113.2M (1.64%)
	ChemDFM-13B	13B (100.%)	13B (100.%)
Full FT	MolT5-Large	0.0M (0.00%)	780.1M (100.%)
	MolT5-Large + MoMu	2.0M (0.00%)	782.1M (100.%)

space using a versatile pre-training strategy. Experimental evaluations show that our proposed method consistently outperforms baselines across 4 downstream molecule-related tasks ranging from molecule-to-text translation to retrosynthesis, with and without fine-tuning the LLM, while only introducing a small number of trainable parameters—0.53% and 0.82%, respectively. As shown in Table 5, our proposed method is designed to be more efficient than most baselines while giving superior performances.

7 Acknowledgement

This work was supported by the National Science Foundation (CHE–2202693) through the NSF Center for Computer Assisted Synthesis (C-CAS).

References

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

[2] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712* (2022).

[3] Zachary J Baum, Xiang Yu, Philippe Y Ayala, Yanan Zhao, Steven P Watkins, and Qiongqiong Zhou. 2021. Artificial intelligence in chemistry: current trends and future directions. *Journal of Chemical Information and Modeling* 61, 7 (2021), 3197–3212.

[4] Esben Jannik Bjerrum and Boris Sattarov. 2018. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* 8, 4 (2018), 131.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

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

[7] Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *International Conference on Machine Learning*. PMLR, 6140–6157.

[8] Laurianne David, Amol Thakkar, Rocío Mercado, and Ola Engkvist. 2020. Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics* 12, 1 (2020), 56.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. doi:10.18653/v1/N19-1423

[10] Hideo Doi, Kazuaki Z Takahashi, and Takeshi Aoyagi. 2022. Screening toward the Development of Fingerprints of Atomic Environments Using Bond-Orientational Order Parameters. *ACS omega* 7, 5 (2022), 4606–4613.

[11] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 375–413. doi:10.18653/v1/2022.emnlp-main.26

[12] Anna Gaulton, Louisa J Bellis, A Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, et al. 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40, D1 (2012), D1100–D1107.

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

[14] Zhichun Guo, Wenhao Yu, Chuxu Zhang, Meng Jiang, and Nitesh V Chawla. 2020. Graseq: graph and sequence fusion learning for molecular property prediction. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 435–443.

[15] Jiazhen He, Huifang You, Emil Sandström, Eva Nittinger, Esben Jannik Bjerrum, Christian Tyrchan, Werngard Czechtizky, and Ola Engkvist. 2021. Molecular optimization by capturing chemist's intuition using deep neural networks. *Journal of cheminformatics* 13 (2021), 1–17.

[16] Yifan Hou, Jian Zhang, James Cheng, Kaili Ma, Richard T. B. Ma, Hongzhi Chen, and Ming-Chang Yang. 2020. Measuring and Improving the Use of Graph Information in Graph Neural Networks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkeIIkHKvS

[17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[18] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2022. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* 3, 1 (2022), 015022.

[19] Jeonghee Jo, Bumju Kwak, Hyun-Soo Choi, and Sungroh Yoon. 2020. The message passing neural networks for chemical property prediction on SMILES. *Methods*

179 (2020), 65–72.

[20] Greg Landrum et al. 2013. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8, 31.10 (2013), 5281.

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[22] Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. Towards 3D Molecule-Text Interpretation in Language Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=xI4yNlkaqh

[23] Jianping Liu, Xiujuan Lei, Yuchen Zhang, and Yi Pan. 2023. The prediction of molecular toxicity based on BiGRU and GraphSAGE. *Computers in biology and medicine* 153 (2023), 106524.

[24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[25] Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling* 62, 6 (2022), 1376–1387.

[26] Harry L Morgan. 1965. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation* 5, 2 (1965), 107–113.

[27] João CA Oliveira, Johanna Frey, Shuo-Qing Zhang, Li-Cheng Xu, Xin Li, Shu-Wen Li, Xin Hong, and Lutz Ackermann. 2022. When machine learning meets molecular synthesis. *Trends in Chemistry* 4, 10 (2022), 863–885.

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[29] Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards Generalized Biological Understanding with IUPAC Integration and Multi-task Tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 1216–1240. https://aclanthology.org/2024.findings-acl.71

[30] Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching Cross-modal Integration in Biology with Chemical Knowledge and Natural Language Associations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 1102–1123. doi:10.18653/v1/2023.emnlp-main.70

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

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

[33] Tatsuya Sagawa and Ryosuke Kojima. 2023. ReactionT5: a large-scale pre-trained model towards application of limited reaction data. *arXiv preprint arXiv:2311.06708* (2023).

[34] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=9Vrb9D0WI4

[35] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling* 56, 12 (2016), 2336–2346.

[36] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1715–1725. doi:10.18653/v1/P16-1162

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

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

[39] Amalio Telenti, Michael Auli, Brian L Hie, Cyrus Maher, Suchi Saria, and John PA Ioannidis. 2024. Large language models for science and medicine. *European Journal of Clinical Investigation* 54, 6 (2024), e14183.

[40] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[41] Varvara Voinarovska, Mikhail Kabeshov, Dmytro Dudenko, Samuel Genheden, and Igor V Tetko. 2023. When yield prediction does not yield prediction: an overview of the current challenges. *Journal of Chemical Information and Modeling* 64, 1 (2023), 42–56.

[42] Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D. Burke. 2022. Chemical-Reaction-Aware Molecule Representation Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=6sh3pIzKS-

[43] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. 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*. 429–436.

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

[45] David Weininger. 1988. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 28, 1 (1988), 31–36.

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

[47] Jun Xia, Lecheng Zhang, Xiao Zhu, Yue Liu, Zhangyang Gao, Bozhen Hu, Cheng Tan, Jiangbin Zheng, Siyuan Li, and Stan Z Li. 2024. Understanding the limitations of deep models for molecular property prediction: Insights and solutions. *Advances in Neural Information Processing Systems* 36 (2024).

[48] Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-BERT: Rethinking Pre-training Graph Neural Networks for Molecules. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=jevY-DtiZTR

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

[50] Botao Yu, Frazier N Baker, Ziqi Chen, Xia Ning, and Huan Sun. 2024. Llasmol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. *arXiv preprint arXiv:2402.09391* (2024).

[51] Lawrence Zhao, Carl Edwards, and Heng Ji. 2023. What a Scientific Language Model Knows and Doesn't Know about Chemistry. In *NeurIPS 2023 AI for Science Workshop*.

[52] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. 2024. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818* (2024).

[53] Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. 2022. Root-aligned SMILES: a tight representation for chemical reaction prediction. *Chemical Science* 13, 31 (2022), 9023–9034.

A Pre-training Strategy

Here we elaborate the pre-training strategy by describing all proposed pre-training tasks. The molecule description generation task serves as the main task, accompanied by a couple of auxiliary tasks. We select a set of 10 low-level properties that present comprehensive information about the molecules. We use one more special auxiliary task which is canonicalizing the molecule's SMILES string. Examples of these tasks and their instructions are illustrated in Figure 7.

B Experimental Settings

The MolX-enhanced LLM is pre-trained with the above tasks in a multi-task learning setting for 5 epochs. AdamW optimizer [24] is adopted with a weight decay of 0.05 and a learning rate scheduler of a combination of linear warmup with 1000 steps and cosine decay, in which the peak and minimal learning rates are 1e-5 and 5e-6,

Molecule description generation

Provide a detailed description of the molecule.
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: The molecule is a cyclo(tyrosyl-tyrosyl) in which both stereocentres have L-configuration. Synthesized by Mycobacterium tuberculosis. It has a role as a metabolite. **100%**

Predicting the basic chemical and physical properties

Heavy Atom

A heavy atom refers to any atom that is not hydrogen.
How many heavy atoms are there in this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 24 **10%**

Hydrogen Bond Acceptor

A hydrogen bond acceptor has lone electron pairs that help form hydrogen bonds.
How many hydrogen bond acceptors are there in this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 4 **10%**

Hydrogen Bond Donor

A hydrogen bond donor is a compound that donates protons (hydrogen atoms) covalently bound to itself, allowing it to form hydrogen bonds.
How many hydrogen bond donors are there in this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 4 **10%**

Rotatable Bond

A rotatable bond is any single non-ring bond, attached to a non-terminal, non-hydrogen atom.
How many rotatable bonds are there in this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 4 **10%**

Aromatic Ring

Aromatic rings are hydrocarbons with a benzene or related ring.
How many aromatic rings are there in this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 2 **10%**

Complexity

The complexity rating of a compound estimates its structural complexity based on its elements and structural features, including symmetry.
What is the complexity rating of this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 412 **10%**

Topological Polar Surface Area

The topological polar surface area (TPSA) is the surface sum of all polar atoms or molecules, primarily oxygen and nitrogen, also including their attached hydrogen atoms.
What is the TPSA value of this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 98.66 **10%**

Molecular weight

The molecular weight is the sum of the atomic weights of all the atoms in the molecule.
What is the molecular weight of this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 326.352 **10%**

LogP

LogP, or octanol-water partition coefficient, is a measure of how hydrophilic or hydrophobic a molecule is.
What is the LogP value of this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 0.8662 **10%**

Quantitative Estimate of Druglikeness

The quantitative estimate of druglikeness (QED) is a measure of how drug-like a molecule is, based on various molecular properties associated with druglikeness.
What is the QED value of this molecule?
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: 0.669781 **10%**

Canonicalizing the molecule's SMILES string

Provide the molecule's canonical SMILES string, which is a unique representation of this molecule.
Molecule: C1=CC(=CC=C1C[C@H]2C(=O)N[C@H](C(=O)N2)CC3=CC=C(C=C3)O)O
Answer: O=C1N[C@@H](Cc2ccc(O)cc2)C(=O)N[C@H]1Cc1ccc(O)cc1 **10%**

Figure 7: Examples of all pre-training tasks in our instruction-based pre-training strategy.

Table 6: Added results for molecule description generation.

	Model	# Trainable Params		Description Generation					
		Pre-training	Downstream	BLE-2↑	BLE-4↑	ROG-1↑	ROG-2↑	ROG-L↑	MET↑
LoRA FT	Llama-2-7B + MolX w/o ChemInit	36.1M (0.53%)	56.6M (0.82%)	30.21	22.67	43.64	28.80	38.47	38.43
	Llama-2-7B + MolX w/o `MorganFP`	23.5M (0.35%)	44.0M (0.64%)	29.33	22.01	42.37	27.96	37.35	37.31
	Llama-2-7B + MolX w/o WeightedInc	36.1M (0.53%)	56.6M (0.82%)	31.13	24.01	44.16	28.50	38.56	39.34
	Llama-2-7B + MolX w/o Auxiliaries	36.1M (0.53%)	56.6M (0.82%)	30.71	23.06	40.29	24.33	33.62	35.37
	Llama-2-7B + MolX w/o Pre-training	00.0M (0.00%)	56.6M (0.82%)	28.79	22.36	38.23	22.28	30.40	33.13
	Llama-2-7B + **MolX**	36.1M (0.53%)	56.6M (0.82%)	**31.40**	**24.25**	**44.20**	**28.96**	**38.76**	**39.55**

respectively. The batch size is 12 and the maximal text length is set to be 256. The computation time is 72 hours on 2 A100 GPUs with BFloat16 mixed precision. For experiments on downstream tasks, we consider two different model configurations for the evaluation: I) Inference-only: The model is frozen after pre-training for direct question answering on downstream tasks, evaluating the model's generalizability without fine-tuning; II) LoRA fine-tuning:

The model is fine-tuned on downstream tasks using a parameter-efficient technique, LoRA [17], verifying the model's adaptability in scenarios where downstream data are available. For LoRA fine-tuning, the model is fine-tuned on train sets of downstream tasks for 50 epochs, using the same settings of optimizer and learning rate scheduler as pre-training. LoRA is applied with the same hyper-parameters as the baselines 2D and 3D MoLM [22], factorizing all $*\_proj$ modules of `LlamaSdpaAttention` and `LlamaMLP` layers with a rank $r = 8$, $\alpha = 32$, and $dropout = 0.1$. Notably, for all tasks, the loss function employed is the auto-regressive loss as described in Equation (4). We report performances on the test sets selected by the corresponding validation sets.

C Ablation Study

Here we study the influence of building components in our proposed framework. Firstly, we use random initializations for trainable encoders, exploring the possibility of eliminating reliance on robust pre-trained weights. Next, we investigate the contributions of incorporating the Morgan fingerprint, as well as the weighted scheme by removing them from the framework. Moreover, to demonstrate the effectiveness of our versatile pre-training strategy, we discard auxiliary tasks and only use the molecule description generation objective during pre-training. Lastly, by totally skipping the pre-training stage, we aim to understand its alignment impact on the framework. Experiments are conducted on the molecule description generation task on the PubChem dataset [22] under the LoRA fine-tuning scenario, simultaneously highlighting the proposed framework's efficiency regarding the number of trainable parameters during pre-training and fine-tuning on downstream tasks.

Table 6 shows experimental results for the described ablation study. Firstly, a drop in the performances of MolX without chemical initializations for encoders indicates the role of robust pre-trained weights. Next, while the weighted scheme brings a modest improvement, incorporating the Morgan fingerprint contributed significantly to the performances of MolX. Moreover, without proposed auxiliary tasks, a noticeable decrease in performances can be viewed, especially for ROUGE-based and METEOR metrics, demonstrating their effectiveness in providing the model with comprehensive information about the molecules. Lastly, it is not surprising that the pre-training stage which forms an alignment between MolX and the LLMs' textual input space, has a large impact. In terms of efficiency, our proposed framework only introduces a small number of trainable parameters, accounting for 0.53% of the entire parameters during pre-training and 0.82% with fine-tuning on downstream tasks.

Received June 2025; revised June 2025; accepted June 2025