Property Enhanced Instruction Tuning for Multi-Task Molecule Generation with Large Language Models

Xuan Lin[1], **Long Chen**[1], **Yile Wang**[2*], **Yangyang Chen**[3] and **Xiangxiang Zeng**[4]
[1]School of Computer Science, Xiangtan University
[2]College of Computer Science and Software Engineering, Shenzhen University
[3] Department of Computer Science, University of Tsukuba
[4] College of Computer Science and Electronic Engineering, Hunan University
wangyile@szu.edu.cn

Abstract

Large language models (LLMs) are widely applied in various natural language processing tasks such as question answering and machine translation. However, due to the lack of labeled data and the difficulty of manual annotation for biochemical properties, the performance for molecule generation tasks is still limited, especially for tasks involving multi-properties constraints. In this work, we present a two-step framework PEIT (**P**roperty **E**nhanced **I**nstruction **T**uning) to improve LLMs for molecular-related tasks. In the first step, we use textual descriptions, SMILES, and biochemical properties as multimodal inputs to pre-train a model called PEIT-GEN, by aligning multi-modal representations to synthesize instruction data. In the second step, we fine-tune existing open-source LLMs with the synthesized data, the resulting PEIT-LLM can handle molecule captioning, text-based molecule generation, molecular property prediction, and our newly proposed multi-constraint molecule generation tasks. Experimental results show that our pre-trained PEIT-GEN outperforms MolT5, BioT5, MolCA and Text+Chem-T5 in molecule captioning, demonstrating modalities align well between textual descriptions, structures, and biochemical properties. Furthermore, PEIT-LLM shows promising improvements in multi-task molecule generation, demonstrating the effectiveness of the PEIT framework for molecular tasks. The code and appendix are available at https://github.com/chenlong164/PEIT.

1 Introduction

Large language models (LLMs) such as GPT-4 [OpenAI, 2023], Gemini [Comanici *et al.*, 2025] and DeepSeek [Guo *et al.*, 2025] have revolutionized the landscape of artificial intelligence and natural language processing, allowing machines to understand and generate human language with remarkable fluency and coherence. Based on encoded world knowledge [Petroni *et al.*, 2019] and powerful instruct-following [Zhang *et al.*, 2023] capabilities of LLMs, recent work has successfully used LLM for molecular-related tasks, achieving promising results [Cao *et al.*, 2025].

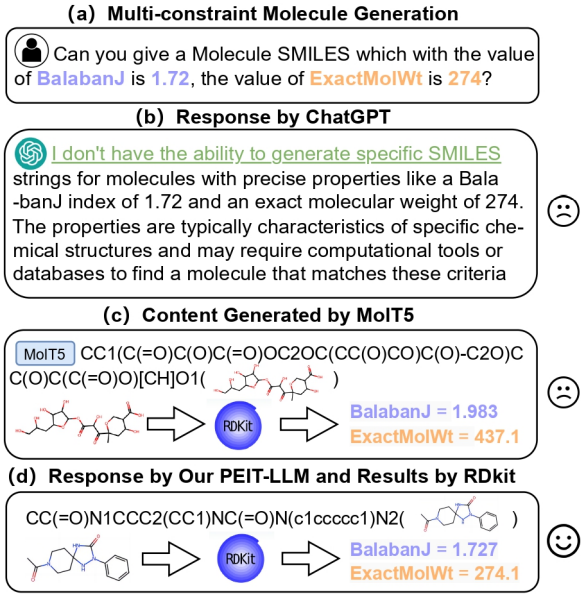

Figure 1: (a) An example of our proposed multi-constraint molecule generation task. (b) The response by ChatGPT. (c) The result generated by MolT5. (d) The response generated by the LLaMA3.1 model after applying our proposed property-enhanced instruction tuning, with the results validated by RDKit.

Despite their success, as illustrated in Figure 1 (a) and (b), LLMs still struggle with generating molecules under strict property constraints. Even specialized molecular translation models, such as MolT5 [Edwards *et al.*, 2022], fail in these tasks (see Figure 1 (c)). While these models effectively capture relationships between molecular text and structure, they lack sufficient understanding of molecular properties, limiting their ability to incorporate property constraints in prompts. This shortcoming restricts their practical utility in applications like drug discovery [Zhavoronkov, 2018].

The challenges in addressing such tasks mainly lie in three aspects: (1) Existing studies have revealed the limitations of

*Corresponding author.

LLMs in understanding molecular representations [Grisoni, 2023], making it difficult to handle tasks requiring precise molecular property comprehension; (2) While there are some known SMILES-property pairing data, it often remains limited to predicting a single property and lacks datasets that cover a wide range of properties [Wu *et al.*, 2018]. Moreover, most of these datasets do not include precisely described textual data, making it challenging to identify accurate trimodal data pairs; (3) To our knowledge, there are no suitable datasets or evaluation methods exist for multi-constraint molecule generation using LLMs, which challenges the standardization and assessment of such tasks [Elton *et al.*, 2019].

To address these challenges, we propose a framework called PEIT (**P**roperty **E**nhanced **I**nstruction **T**uning) to generate multi-modal molecular instruction datasets in bulk, aiming to enhance capabilities of LLMs in multi-task molecule generation. Using the PEIT framework, our pre-trained model can handle general tasks (e.g., molecule captioning [Edwards *et al.*, 2022]) and property-related tasks such as property prediction [Chang and Ye, 2024]. This makes it suitable for constructing data to evaluate multi-constraint molecule generation capabilities and for serving as instruction tuning data to improve existing open-source LLMs.

The overall structure of the proposed PEIT framework is shown in the left of Figure 2. Specifically, it consists of two components: (1) We pre-train a model called PEIT-GEN through multi-modal representation alignment, which integrates text-based (molecular descriptions), structure-based (SMILES), and property-based (property-value pairs) information to generate diverse unstructured text, sequence, and property data; (2) By using the synthesized instruction data, we fine-tune open-source LLMs and develop PEIT-LLM, which can be applied to various molecule generation tasks mentioned above, including our proposed multi-constraint molecule generation.

Experimental results demonstrate that our pre-trained PEIT-GEN achieves competitive or better results in molecule captioning tasks, comparing to a variety of biomolecular models including MolT5 [Edwards *et al.*, 2022], BioT5 [Pei *et al.*, 2023], GIT-Mol [Liu *et al.*, 2024], MolXPT [Liu *et al.*, 2023a], MolCA [Liu *et al.*, 2023b], and Text+Chem-T5 [Christofidellis *et al.*, 2023]. Additionally, PEIT-LLM based on LLaMa3.1-8B [Dubey *et al.*, 2024] exhibits superior performance compared to specialized models Mol-Instructions [Fang *et al.*, 2023] and general-purpose LLMs including LLaMa3 [Dubey *et al.*, 2024] and Qwen2.5 [Yang *et al.*, 2024] in molecular property prediction and our newly proposed multi-constraint molecule generation tasks.

2 Related Work

Molecule Generation. Molecule generation tasks fall into two categories: (1) text-based molecule generation that uses textual descriptions to generate molecules that match the given description [Liu *et al.*, 2023a; Liu *et al.*, 2024]. MolT5 [Edwards *et al.*, 2022] was the first proposed to realize translation between textual description and molecular SMILES. BioT5 aims to enhance molecular understanding by incorporating protein. They also perform molecule captioning, which is equivalent to the inverse task of text-based molecule generation. (2) property-guided molecule generation is the inverse process of molecular property prediction, where molecules are generated based on specific biochemical property constraints. Notably, SPMM [Chang and Ye, 2024] was the first to establish a connection between 53 biochemical properties and SMILES sequences, making multi-constraint molecule generation possible. However, few existing models can simultaneously perform text-based or multi-constraint molecule generation and molecule captioning.

Molecular Property Prediction. Deep learning models have been developed for molecular property prediction each with their own advantages and limitations. Transformer-based models design attention mechanism to capture contextual contexts from large-scale SMILES sequences [Ross *et al.*, 2022]. The molecular graph can be directly obtained from SMILES sequences via RDKit [Landrum and others, 2013]. Graph-based models develop diverse graph neural networks to learn differentiable representations [Wang *et al.*, 2022]. However, these methods ignore the potential that incorporating textual knowledge enables to realize new drug design objectives. Recently, a novel molecular pre-trained model named SPMM [Chang and Ye, 2024] that extends the application of multimodal pre-training approaches by aligning molecular structures and biochemical properties. This paper extends the multimodal pre-training to patterns of text-sequence-property triplets, which is defined flexibly by LLM-understandable textual prompts.

Instruction Tuning. Constructing specialized instruction datasets is an effective way to enable LLMs to better perform molecular-related tasks. For instance, Mol-Instructions [Fang *et al.*, 2023] provides a large-scale biomolecular dataset tailored for LLMs, covering a variety of instructions involving small molecules, proteins, and biomolecular texts. ChemDFM [Zhao *et al.*, 2025] advances this paradigm by creating a broader dataset spanning molecular structures, reactions, and properties. Its two-stage training—domain pretraining followed by instruction tuning—enhances the model's chemical understanding and reasoning capabilities. More recently, GeLLM3O [Dey *et al.*, 2025] introduced MuMOInstruct, a high-quality dataset focused on multi-property molecule optimization, demonstrating strong generalization across diverse tasks. Despite these advances, generating reliable and scalable molecular instruction data remains a key challenge, particularly for open-source models.

3 Method

3.1 Overview of PEIT Framework

The overview of PEIT framework is shown in Figure 2 (left), which consists of PEIT-GEN and PEIT-LLM. In PEIT-GEN, we generate a large number of "SMILES-text" and "SMILES-property" pairs to serve as multi-modal data. Then we design multiple multi-modal alignment objectives to pre-train PEIT-GEN. In PEIT-LLM, by using the pre-trained PEIT-GEN, we can predict a large number of triplets to generate more diverse SMILES inputs, and then construct diverse instruction data based on template filling. By utilizing the synthesized instruction data, PEIT-LLM enables

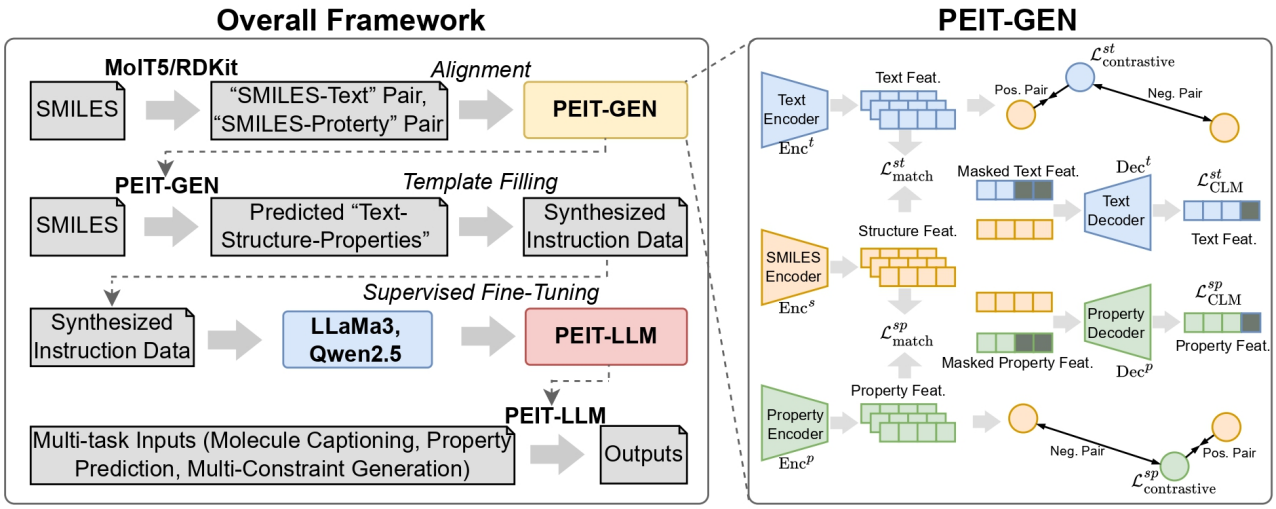

Figure 2: Left: Overall PEIT framework. We first pre-train the PEIT-GEN and construct instruction data via template filling. Then we fine-tune the open-source LLMs through instruction tuning, the resulting PEIT-LLM is used for multi-task molecule generation. Right: The process of PEIT-GEN pre-training, see details in Section 3.2.

the supervised fine-tuning of open-source LLMs including LLaMa [Dubey *et al.*, 2024] and Qwen [Yang *et al.*, 2024], enhancing the capabilities for multi-task molecule generation.

3.2 Pre-training of PEIT-GEN

The pre-training stage of PEIT-GEN is shown in the right of Figure 2. For a given molecule, different representations offer unique and complementary features, which are crucial for comprehensive molecule understanding. PEIT-GEN aims to integrate information from three modalities simultaneously, including textual information $\mathcal{T}$ (text), molecular structure $\mathcal{S}$ (SMILES), and biochemical properties $\mathcal{P}$ (property-value). Such ability can help synthesizing sufficient instruction data for further enhancing the ability of LLMs. In particular, PEIT-GEN consists of three Transformer encoders Enc^t, Enc^s, Enc^p and two decoders Dec^t, Dec^p, and we design different training objectives to align features from different modalities.

Cross-modal Representation Matching. Following SPMM [Chang and Ye, 2024], we leverage pre-trained models SciBERT [Beltagy *et al.*, 2019] as trainable Enc^t for encoding textual data, BERT [Devlin *et al.*, 2019] as Enc^s and Enc^p for encoding SMILES and properties. Then we obtain feature representations across all three modalities, establishing the foundation for feature alignment.

We propose cross-modal representation matching to align the representations from different perspectives by the same molecule. In particular, we introduce the SMILES-text matching loss $\mathcal{L}^{st}_{\text{match}}$ and the SMILES-property matching loss $\mathcal{L}^{sp}_{\text{match}}$, which serve as objectives for training the encoders. In this way, the model can effectively learn cross-modal relationships and improve performance in multi-modal tasks by aligning the feature spaces. The matching loss is calculated as follows:

$$\mathcal{L}^{st}_{\text{match}} = \ell_{\text{CE}}\left(y^{st}_{\text{match}}, \text{MLP}(\text{Enc}^s(\mathcal{S}) \oplus \text{Enc}^t(\mathcal{T}))\right), \quad (1)$$

$$\mathcal{L}^{sp}_{\text{match}} = \ell_{\text{CE}}\left(y^{sp}_{\text{match}}, \text{MLP}(\text{Enc}^s(\mathcal{S}) \oplus \text{Enc}^p(\mathcal{P}))\right), \quad (2)$$

where y^{st}_{match} and y^{sp}_{match} are labels as 0 or 1, indicating whether the corresponding SMILES-text or SMILES-property pairs are matching. $\text{Enc}(\cdot)$ indicates the representation of the data (i.e., [CLS] token of Transformer encoder), $\oplus$ is the concatenation operation, and $\text{MLP}(\cdot)$ is the trainable multi-layer perception. The encoders are optimized by the cross-entropy loss ℓ_{CE} using the given data from different modalities.

Multi-modal Contrastive Learning. The representation matching can be viewed as an explicit 2-way classification training. We further utilize contrastive learning to directly enhancing the representation by pulling semantically close neighbors together and pushing apart non-neighbors from data of different modalities. To calculate the similarity between the encoded features of different modalities, we extract the encoded features and then compute the instance-level similarities through the inner product:

$$\text{sim}(\mathcal{S}, \mathcal{T}) = (\text{MLP}^s(\text{Enc}^s(\mathcal{S})))^\top \text{MLP}^t(\text{Enc}^t(\mathcal{T})), \quad (3)$$

$$\text{sim}(\mathcal{S}, \mathcal{P}) = (\text{MLP}^s(\text{Enc}^s(\mathcal{S})))^\top \text{MLP}^p(\text{Enc}^p(\mathcal{P})), \quad (4)$$

where MLP^s, MLP^t and MLP^p are multi-layer perceptions applied to SMILES, text, and property representations, respectively. Then, for the given SMILES $\mathcal{S}$, text $\mathcal{T}$, and property $\mathcal{P}$, we compute the cross-modal batch-level similarities as follows:

$$s_{s2t} = \frac{\exp(\text{sim}(\mathcal{S},\mathcal{T})/\tau)}{\sum_{i=1}^{M}\exp(\text{sim}(\mathcal{S},\mathcal{T}_i)/\tau)}, \quad (5)$$

$$s_{s2p} = \frac{\exp(\text{sim}(\mathcal{S},\mathcal{P})/\tau)}{\sum_{i=1}^{N}\exp(\text{sim}(\mathcal{S},\mathcal{P}_i)/\tau)}, \quad (6)$$

where M and N represent the total number of texts and property in the batch of data pairs, respectively. τ is the temperature controlling the sharpness of the similarity. The intra-modal similarities s_{s2s}, s_{p2p}, and s_{t2t} can be computed in similar manners.

Based on the cross-modal and intra-modal batch-level similarities, the contrastive loss is formulated by calculating the

cross-entropy according to one-hot encoded similarity vectors y, where the value is 1 for pairs derived from the same molecule or 0 for all other combinations:

$$\mathcal{L}^{st}_{\text{contrastive}} = \frac{1}{2}(\ell_{\text{CE}}(y_{s2t}, s_{s2t}) + \ell_{\text{CE}}(y_{t2s}, s_{t2s}) + \ell_{\text{CE}}(y_{s2s}, s_{s2s}) + \ell_{\text{CE}}(y_{t2t}, s_{t2t})), \quad (7)$$

$$\mathcal{L}^{sp}_{\text{contrastive}} = \frac{1}{2}(\ell_{\text{CE}}(y_{s2p}, s_{s2p}) + \ell_{\text{CE}}(y_{p2s}, s_{p2s}) + \ell_{\text{CE}}(y_{s2s}, s_{s2s}) + \ell_{\text{CE}}(y_{p2p}, s_{p2p})). \quad (8)$$

Cross-modal Causal Language Modeling. To further strengthen the model's capability in molecule captioning, we employ the causal language modeling (CLM) to enhance the model performance on text generation. Specifically, we design decoders to generate subsequent property and textual description sequences, under the guidance of SMILES features through cross-attention as show in Figure 3.

By introducing the SMILES features in attention layers for CLM training, the cross-modal CLM loss $\mathcal{L}^{st}_{\text{CLM}}$ and $\mathcal{L}^{sp}_{\text{CLM}}$ are computed as follows:

$$\mathcal{L}^{st}_{\text{CLM}} = -\sum_{i=1}^{N}\sum_{j=1}^{n} \log \text{Prob}\left(w_j^{(i)} \mid \text{Dec}^t(\tilde{\mathbf{w}}_{:j}^{(i)}); \theta_t\right), \quad (9)$$

$$\mathcal{L}^{sp}_{\text{CLM}} = -\sum_{i=1}^{N}\sum_{j=1}^{n} \log \text{Prob}\left(w_j^{(i)} \mid \text{Dec}^p(\tilde{\mathbf{w}}_{:j}^{(i)}); \theta_p\right), \quad (10)$$

where Prob is the conditional probability to predict the word $w_j^{(i)}$ in the vocabulary, N is the total number of samples, n is the index of current words in each sample, $\tilde{\mathbf{w}}_{:j}^{(i)}$ is the sequence from begin to the j-th word in the i-th sample, θ_t and θ_p are the trainable parameters in two decoders.

Training. The overall training objective for pre-training PEIT-GEN is to minimize the sum of all three types of losses across three modalities:

$$\mathcal{L} = \mathcal{L}^{st}_{\text{match}} + \mathcal{L}^{sp}_{\text{match}} + \alpha\mathcal{L}^{st}_{\text{contrastive}} + \alpha\mathcal{L}^{sp}_{\text{contrastive}} + \beta\mathcal{L}^{st}_{\text{CLM}} + \beta\mathcal{L}^{sp}_{\text{CLM}}, \quad (11)$$

where we conducted a hyperparameter search for α and β, as shown in Table 1 in Appendix A. Based on the experimental results, we follow the setting in SPMM[Chang and Ye, 2024] and adopt a ratio of $\alpha : \beta = 1 : 5$ to balance the loss terms.

3.3 Instruction Tuning for PEIT-LLM

Template Filling. The pre-trained PEIT-GEN offers unstructured data in the format of "text-SMILES-properties" (i.e., text-structure-property) triplets, which are stored in CSV files containing text, molecular structures, and information on 53 molecular biochemical properties. To obtain more task-specific data and to adapt to the strong instruction-following abilities of LLMs, we design templates for different downstream tasks, as shown in Figure 1 of Appendix B. For text-based molecule generation as example, we fix a general question format and then extract molecular descriptions from unstructured data to fill the pre-defined template, resulting in a natural question as instructions. The SMILES from unstructured triplets is used as the desired response. In this way, we can generate diverse task-specific instruction data in bulk for subsequent instruction-tuning.

Model	MC	TBMG	MPP	MCMG
MolT5	✓	✓	✗	✗
BioT5	✓	✓	✗	✗
MolXPT	✓	✓	✗	✗
Git-Mol	✓	✓	✗	✗
SPMM	✗	✗	✓	✗
MolCA	✓	✓	✗	✗
Text+Chem-T5	✓	✓	✗	✗
BioMedGPT	✓	✗	✗	✗
InstructMol-GS	✓	✗	✗	✗
MolReGPT	✓	✓	✗	✗
Mol-Instructions	✓	✓	✓ (poor)	✓ (poor)
LLaMa, Qwen	✓ (limited)	✓ (poor)	✓ (poor)	✓ (poor)
PEIT-LLM (Ours)	✓	✓	✓	✓

Table 1: Comparing PEIT-LLM with biomolecular models and LLMs on molecular-related tasks. MC: Molecule Captioning. TBMG: Text-Based Molecule Generation. MPP: Molecular Property Prediction. MCMG: Multi-Constraint Molecule Generation.

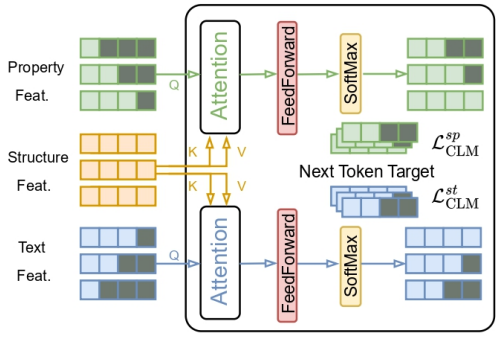

Figure 3: The cross-modal causal language modeling.

Multi-constraint Molecule Generation Task. Molecule generation often requires to be conducted under multiple constraints rather than a single condition. In this work, we propose a new task to assess molecule generation through a variety of descriptors, by comparing the alignment between the generated molecules and specific criteria to evaluate the generative performance of LLMs. By using the large-scale unstructured data generated by PEIT-GEN, we can effectively synthesize sufficient data for evaluation. Specifically, we follow SPMM [Chang and Ye, 2024] given the vast number of molecular attributes and the complex combinations thereof, analyzing their impact on results across different molecular counts poses a signifi- cant computational challenge due to resource limitations. To address this, we selected representative ADMET [Fu *et al.*, 2024] properties—BalabanJ, MolLogP, ExactMolWt, QED, and TPSA—that capture molecular topology, electronic characteristics, and steric effects, while exhibiting low mutual correlation. Based on template filling, the predicted multi-property values are used to construct data for multi-constraint molecule generation. We employ instruction tuning to guide the LLM in generating molecules, and use RDKit [Landrum and others, 2013] to calculate the actual property values. RMSE and R^2 are then used to compare these values against the constraints, enabling a systematic evaluation of the LLM's performance in multi-constraint molecule generation tasks.

Supervised Fine-tuning. We select LLaMa3.1-8B [Dubey *et al.*, 2024] and Qwen2.5-7B [Yang *et al.*, 2024] as base

Model	Data Size ↓	BLEU-2 ↑	BLEU-4 ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
MolT5-small [Edwards *et al.*, 2022]	100M	0.513	0.398	0.492	0.567	0.412	0.501
MolT5-large [Edwards *et al.*, 2022]	100M	0.594	0.508	0.613	0.654	0.508	0.592
BioT5 [Pei *et al.*, 2023]	33M	**0.635**	**0.556**	0.656	0.692	0.559	0.633
MolXPT [Liu *et al.*, 2023a]†	30M	0.594	0.505	0.626	0.660	0.511	0.597
$\text{MolCA}_{\text{w/ Galac}}$ [Liu *et al.*, 2023b]	2.3M	0.616	0.524	0.639	0.674	0.533	0.615
$\text{Text+Chem-T5}_{\text{augm}}$ [Christofidellis *et al.*, 2023]	11.5M	0.625	0.529	0.648	0.682	0.543	0.622
GIT-Mol [Liu *et al.*, 2024]†	4.8M	0.352	0.263	0.533	0.575	0.485	0.560
PEIT-GEN (Ours)	**0.48M**	0.598	0.534	**0.676**	**0.700**	**0.582**	**0.653**

Table 2: Results on CHEBI-20 dataset for molecule captioning with different pre-trained models. †: Reported from papers accordingly. The best results in each column are **in bold**, and the second-best results are underlined.

LLMs. We then perform standard SFT [Ouyang *et al.*, 2024] by using the "instruction-response" pairs. In practice, we construct totally 1 million instruction data of four different tasks (i.e., molecule captioning, text-based molecule generation, property prediction, and multi-constraint molecule generation) from 200k unstructured "text-SMILES-properties" triplets obtained by PEIT-GEN.

3.4 Comparing PEIT-LLM with Biomolecular Models and Large Language Models

Table 1 shows a comparison of our PEIT-LLM with existing pre-trained models and general LLMs on multiple molecular generation tasks. For most of the pre-trained models such as MolT5 and BioT5, they focus on molecule captioning and text-based molecule generation, which can not handle property-related tasks. SPMM is a specialized model for property prediction. However, it lacks generation ability due to the lack of textual descriptions. Current LLMs such as LLaMa and Qwen show strong performance on general NLP-based tasks through conversations or instruction-following. However, these general LLMs still have limitations in tasks related to molecule generation due to a lack of molecular knowledge. In contrast, through fine-tuning on diverse instruction data with rich molecular knowledge, PEIT-LLM can perform multiple molecule generation tasks simultaneously.

4 Experiments

4.1 Experimental Setup

Dataset. During the pre-training of PEIT-GEN, we extract 480k molecular SMILES from the ZINC dataset [Irwin *et al.*, 2012] and generate corresponding SMILES–text pairs using MolT5. Meanwhile, we compute 53 biochemical properties for each molecule using RDKit, forming 480k "text–SMILES–property" triplets for training. We then further fine-tune the model on the CHEBI-20 [Edwards *et al.*, 2021] dataset to enhance its performance on human-annotated data.

For pre-training PEIT-LLM, we utilize the 200k tri-modal data generated by PEIT-GEN and employ template filling to generate 200k instruction data for each downstream task. For molecular property prediction, we select two biochemical properties with distinct differences for evaluation, generating 200k instruction data for each property. Finally, we obtain a total of 1000k instruction data across four tasks for SFT. Similar to PEIT-GEN, molecular property prediction tasks on PEIT-LLM can be validated by RDKit on CHEBI-20 dataset.

Model	BBBP	BACE	Clintox	SIDER
D-MPNN [Yang and others, 2019]	71.0±0.3	80.9±0.6	90.6±0.6	57.0±0.7
N-GramRF [Liu *et al.*, 2019]	69.7±0.6	77.9±1.5	77.5±4.0	66.8±0.7
N-GramXGB [Liu *et al.*, 2019]	69.1±0.8	79.1±1.3	87.5±2.7	65.5±0.7
PretrainGNN [Hu *et al.*, 2019]	68.7±1.3	84.5±0.7	72.6±1.5	62.7±0.8
$\text{GROVER}_{\text{large}}$ [Rong *et al.*, 2020]	69.5±0.1	81.0±1.4	76.2±3.7	65.4±0.1
ChemRL-GEM [Fang *et al.*, 2022]	72.4±0.4	85.6±1.1	90.1±1.3	**67.2±0.4**
ChemBERTa [Ahmad *et al.*, 2022]†	72.8	79.9	56.3	-
MolFormer [Ross *et al.*, 2022]	73.6±0.8	**86.3±0.6**	91.2±1.4	65.5±0.2
SPMM [Chang and Ye, 2024]	**74.1±0.6**	82.9±0.3	90.7±0.5	63.6±0.5
PEIT-GEN (Ours)	73.6±0.7	81.6±0.5	**91.2±0.7**	62.7±0.9

Table 3: Results on MoleculeNet dataset for 2-way property prediction. †: The standard deviation and the results on SIDER are not reported in literature.

To evaluate PEIT-GEN and PEIT-LLM, we follow MolT5 by using CHEBI-20 and MoleculeNet dataset [Wu *et al.*, 2018], with the standard splition into training, validation, and test sets with an 8:1:1 ratio. All property values are verified via RDKit. See details in Appendix C.

Baseline Models. We compare our model, PEIT-GEN and PEIT-LLM, against three types of baselines as follows: *Baselines on molecule caption* such as MolT5 [Edwards *et al.*, 2022], BioT5 [Pei *et al.*, 2023], MolCA [Liu *et al.*, 2023b], Text+Chem-T5 [Christofidellis *et al.*, 2023], GIT-Mol [Liu *et al.*, 2024]. *Baselines on moleucular property prediction* such as SPMM [Chang and Ye, 2024], D-MPNN [Yang and others, 2019], PretrainGNN [Hu *et al.*, 2019], GROVER [Rong *et al.*, 2020], ChemRL-GEM [Fang *et al.*, 2022]. *Baselines of LLMs* such as LLaMa3 [Touvron *et al.*, 2023], GPT3.5-turbo [OpenAI, 2023], Mol-Instructions [Fang *et al.*, 2023], Qwen2.5 [Yang *et al.*, 2024], BioMedGPT [Zhang *et al.*, 2024], ChatGLM [GLM *et al.*, 2024], LLaSMol [Yu *et al.*, 2024],ChemDFM [Zhao *et al.*, 2025], InstructMol-GS [Cao *et al.*, 2025], Gemini [Comanici *et al.*, 2025], Details and evaluation metric are in Appendix D and E, respectively.

Implementation Details. We pre-train PEIT-GEN for 20 epochs using a batch size of 16, temperature $\tau = 0.07$, and momentum 0.995 with the AdamW optimizer [Loshchilov, 2017]. Fine-tuning is then performed on the CHEBI-20 training set for 200 epochs with a learning rate of 5e-4. For supervised fine-tuning of PEIT-LLM, we use the LLaMa-Factory [Zheng *et al.*, 2024] framework with LoRA [Hu *et al.*, 2022] for 6 epochs, a batch size of 3, and learning rate of 5e-5. The total parameter count of the three encoders and two decoders is 533.8M, with an additional 2.2M parameters for other components. Experiments are conducted on NVIDIA 4090 GPUs with 24GB memory.

Model	#Params	BLEU-2 ↑	BLEU-4 ↑	METEOR ↑	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
LLaMa3 [Touvron *et al.*, 2023]	7B	0.032	0.002	0.117	0.121	0.010	0.065
LLaMa3.1 [Dubey *et al.*, 2024]	8B	0.042	0.004	0.121	0.140	0.019	0.095
Qwen2.5 [Yang *et al.*, 2024]	7B	0.049	0.007	0.188	0.177	0.029	0.112
GPT-3.5-turbo [OpenAI, 2023]	N/A†	0.103	0.050	0.161	0.261	0.088	0.204
Mol-Instructions [Fang *et al.*, 2023]	8B	0.217	0.143	0.254	0.337	0.196	0.291
BioMedGPT [Zhang *et al.*, 2024]	10B	0.234	0.141	0.308	0.386	0.206	0.332
InstructMol-GS [Cao *et al.*, 2025]	7B	0.475	0.371	0.509	0.566	0.394	0.502
MolReGPT [Li *et al.*, 2024]	N/A†	**0.565**	**0.482**	**0.585**	**0.623**	**0.450**	**0.543**
LLaSMol [Yu *et al.*, 2024]	7B	-	-	0.452	-	-	-
ChemDFM [Zhao *et al.*, 2025]	13B	0.321	0.265	0.402	0.490	0.374	0.483
Gemini3 [Comanici *et al.*, 2025]	N/A†	0.141	0.073	0.208	0.184	0.097	0.276
PEIT-LLM-Qwen2.5 (Ours)	7B	0.422	0.314	0.468	0.535	0.361	0.477
PEIT-LLM-LLaMa3.1 (Ours)	8B	0.461	0.356	0.502	0.569	0.396	0.505

Model	#Params	BLEU ↑	Validity ↑	Levenshtein ↓	MACCS FTS ↑	Morgan FTS ↑	RDKit FTS ↑
LLaMa3 [Touvron *et al.*, 2023]	7B	0.261	0.330	45.788	0.372	0.127	0.213
LLaMa3.1 [Dubey *et al.*, 2024]	8B	0.270	0.368	43.183	0.411	0.138	0.248
Qwen2.5 [Yang *et al.*, 2024]	7B	0.217	0.245	50.550	0.403	0.110	0.276
GPT-3.5-turbo [OpenAI, 2023]	N/A†	0.489	0.802	52.130	0.705	0.367	0.462
Mol-Instructions [Fang *et al.*, 2023]	8B	0.345	**1.000**	41.367	0.412	0.147	0.231
MolReGPT [Li *et al.*, 2024]	N/A†	0.790	0.887	24.910	0.847	0.624	0.708
LLaSMol [Yu *et al.*, 2024]	7B	-	-	-	-	0.617	-
Gemini3 [Comanici *et al.*, 2025]	N/A†	0.632	**1.000**	29.782	0.793	0.569	0.638
PEIT-LLM-Qwen2.5 (Ours)	7B	0.810	0.950	21.133	0.832	0.619	0.735
PEIT-LLM-LLaMa3.1 (Ours)	8B	**0.836**	0.970	**18.030**	**0.875**	**0.661**	**0.776**

Table 4: Results on CHEBI-20 dataset for molecule captioning (top) and text-based molecule generation (bottom) tasks. †: MolReGPT is based on closed-source ChatGPT-3.5 and its parameter size remains unknown.

4.2 Comparing PEIT-GEN with Pre-trained Biomolecular Models

Molecule Captioning. Results on molecule captioning using CHEBI-20 dataset are shown in Table 2. Our model demonstrates superior performance in generating high-quality and relevant molecular caption. PEIT-GEN achieved the best results in METEOR and ROUGE, and the second-best performance in BLEU-4. Compared to BioT5 which performs the best in BLEU, our approach requires significantly less data. This indicates that using domain-specific models to generate paired data for pre-training is more efficient than single-modality pre-training.

Molecular Property Prediction. We evaluate the generalization ability of PEIT-GEN on the MoleculeNet benchmark [Wu *et al.*, 2018] using four widely adopted tasks. As shown in Table 3, PEIT-GEN outperforms specialized models such as MolFormer [Ross *et al.*, 2022] and ChemRL-GEM [Fang *et al.*, 2022] on the Clintox dataset. Despite using less pre-training data, it remains competitive on other subsets. To further demonstrate its predictive strength across 53 molecular properties, we present a relative difference analysis in Figure 2 of Appendix F, highlighting PEIT-GEN's strong generalization in property prediction.

4.3 Comparing PEIT-LLM with LLMs

Molecule Captioning. As shown in the top of Table 4, the comparison results show that our model outperforms general-purpose Qwen-2.5 and LLaMa3.1 as well as Mol-Instructions and BioMedGPT, which were trained using a biochemical information instruction dataset for SFT. PEIT-LLM achieved the second-best performance on the ROUGE metric and demonstrated competitive results compared to InstructMol-GS, which was trained solely on the CHEBI-20 dataset and has a similar parameter scale as our base model.

Model	MolWt PP	MolLogP PP	Five-Property CG	
	(RMSE) ↓	(RMSE) ↓	(RMSE) ↓	(R^2) ↑
LLaMa3 [Touvron *et al.*, 2023]	491.542	561.523	79.125	-0.639
LLaMa3.1 [Dubey *et al.*, 2024]	544.517	552.521	74.646	-0.652
Qwen2.5 [Yang *et al.*, 2024]	100.161	132.141	75.991	-0.967
Mol-Instructions [Fang *et al.*, 2023]	72.172	1.313	71.991	-0.352
PEIT-LLM-Qwen2.5 (ours)	14.164	0.164	19.750	0.550
PEIT-LLM-LLaMa3.1 (ours)	**13.918**	**0.141**	**14.212**	**0.613**

Table 5: Results on property prediction (PP), and five-property constraint molecule generation (CG) with different LLMs.

Text-based Molecule Generation. Results on the CHEBI-20 test set are presented at the bottom of Table 4. PEIT-LLM outperforms all baselines on numerical metrics, including BLEU, Levenshtein Distance, and fingerprint similarities based on MACCS, Morgan, and RDKit. Although Mol-Instructions achieves the highest Validity score, the results demonstrate that PEIT-LLM, after multi-task instruction fine-tuning, effectively captures key molecular structures and corresponding textual representations. Case study in Table 5 of Appendix G further supports these findings and indirectly validates the quality of the data generated by PEIT-GEN.

Molecular Property Prediction. For single-property prediction, due to the large number of available properties, we select two representative examples: ExactMolWt, which typically has large numerical values (100 to 1000), and MolLogP, with smaller values (–5 to 10), as shown in Table 5. The results show that PEIT-LLM consistently outperforms other LLMs in predicting these biochemical properties, demonstrating strong sensitivity and adaptability to molecular property scales. This highlights the effectiveness of multi-task

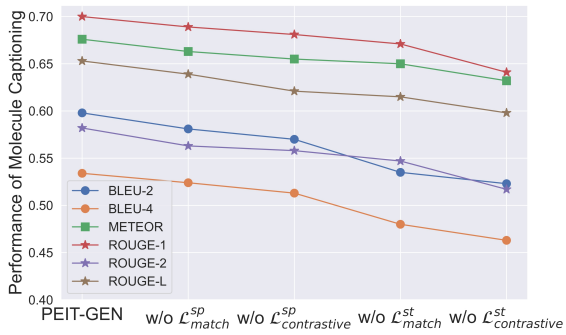

Figure 4: Ablation study on PEIT-GEN pre-training objectives $\mathcal{L}^{sp}_{\text{match}}$, $\mathcal{L}^{st}_{\text{match}}$, $\mathcal{L}^{sp}_{\text{contrastive}}$, and $\mathcal{L}^{st}_{\text{contrastive}}$.

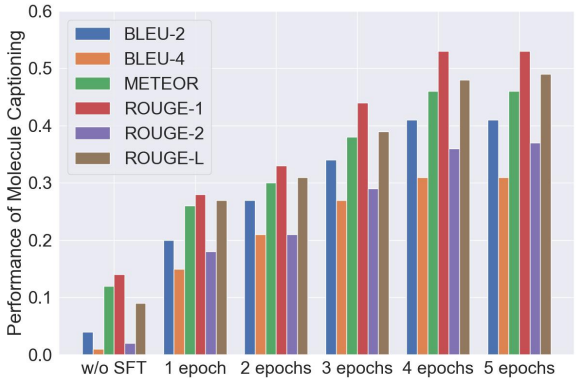

Figure 5: The impact of SFT steps on molecule captioning.

SFT in enhancing LLMs' understanding of molecular characteristics and further validates the quality and reliability of our molecular property instruction dataset. A case study is provided in Table 6 of Appendix G for further illustration.

Multi-constraint Molecule Generation. Results for our proposed multi-constraint molecule generation task is shown in Table 5. PEIT-LLM surpasses baselines by large margin in both RMSE and R^2 metrics. Case study is provided in Table 7 of Appendix G to further illustrate this point. Note that this task requires the model to meet the demands of multiple properties with precise values, placing high demands on the model's overall understanding capability. General-purpose LLMs, or those not specifically trained for this task, lack the required information storage and fitting abilities. The model gain strong molecular understanding capabilities through property enhanced instruction tuning.

4.4 Analyses

Ablation Study. Figure 4 presents an ablation study of the cross-modal matching loss $\mathcal{L}_{\text{match}}$ and cross-modal contrastive loss $\mathcal{L}_{\text{contrastive}}$in the PEIT-GEN model for the molecule captioning task ($\mathcal{L}^{st}_{\text{CLM}}$ and $\mathcal{L}^{sp}_{\text{CLM}}$ are necessary for generation via decoders, thus we do not consider them in ablation study). By removing these training objectives, the performance degradation across all metrics. This demonstrates that both $\mathcal{L}_{\text{match}}$ and $\mathcal{L}_{\text{contrastive}}$ are helpful in cross-modal feature alignment, thereby enhancing the performance

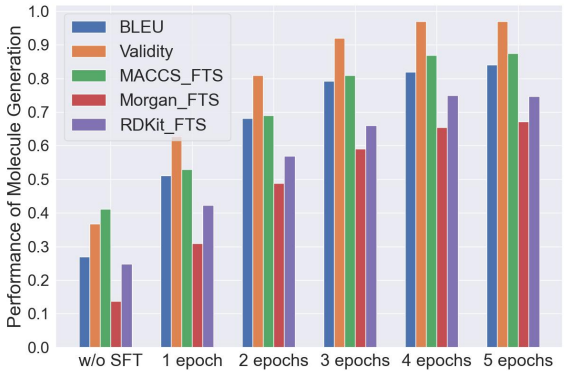

Figure 6: The impact of SFT steps on molecule generation.

Model	BLEU↑	METEOR↑	ROUGE-2↑	ROUGE-L↑
Galactica-6.7B [Taylor *et al.*, 2022]	0.008	0.065	0.015	0.063
MolT5-248M [Edwards *et al.*, 2022]	0.001	0.033	0.001	0.034
Vicuna-7B [Chiang *et al.*, 2023]	0.011	0.168	0.055	0.130
Text+Chem T5-223M [Christofidellis *et al.*, 2023]	0.036	0.139	0.075	0.119
ChatGLM-6B [GLM *et al.*, 2024]	0.011	0.105	0.066	0.148
LLaMa3.1-8B [Dubey *et al.*, 2024]	0.014	0.184	0.066	0.148
Qwen2.5-7B [Yang *et al.*, 2024]	0.009	0.169	0.047	0.119
PEIT-LLM-Qwen2.5-7B (Ours)	0.051	0.208	0.121	0.178
PEIT-LLM-LLaMa3.1-8B (Ours)	0.053	0.215	0.125	0.184
Mol-Instructions-8B [Fang *et al.*, 2023]	0.143	0.254	0.196	0.291

Table 6: Out-of-distribution results on the molecule captioning task using Mol-Instructions [Fang *et al.*, 2023] evaluation set. Mol-Instructions denotes a fully supervised baseline trained with LLaMA3.1-8B using the entire training instructions, serving as an upper bound for SFT models.

of molecule captioning. We further conducted an ablation study on the above loss components in the molecular property prediction task, as shown in the Table 3 of Appendix G.

Impact of SFT steps. Figure 5 and Figure 6 illustrate the outcomes of PEIT-LLM across various tasks with different SFT steps. We observe that the performance consistently improves during the initial epochs for all tasks, indicating that the instructional data is beneficial for each, where the performance tends to plateau around epochs 5-6.

Out-of-Distribution Evaluation. To further evaluate the molecular understanding of PEIT-LLM on unseen data, we tested it on the Mol-Instructions [Fang *et al.*, 2023] test set, without using the full instructions for pre-training and fine-tuning PEIT-LLMs. As shown in Table 6, PEIT-LLM outperforms all general-purpose LLMs as well as smaller domain-specific models highlighting the strong generalization ability of PEIT-LLM across diverse molecular instruction tasks.

5 Conclusion

We propose PEIT, a framework that enables LLMs to perceive multi-modal features for multi-task molecule generation. PEIT aligns molecular structures, textual descriptions, and biochemical properties through multi-modal representation learning. It leverages templates to synthesize diverse, task-specific instruction data for LLMs. We also introduce a challenging multi-constraint molecule generation task, which requires generating novel molecules that satisfy multiple property constraints. Results show that PEIT outperforms various biomolecular models and LLMs on captioning, generation, and property prediction tasks.

Acknowledgments

The work is supported in part by the National Natural Science Foundation of China (62573372, 62425204, U22A2037, 62450002, 62432011), the Hunan Provincial Key Research and Development Program Project (2025JK2003), the Guangdong Basic and Applied Basic Research Foundation (2026A1515011358), the Shenzhen Natural Science Foundation (JCYJ20250604181610014), and the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM602).

References

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

[Beltagy *et al.*, 2019] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3615–3620, 2019.

[Cao *et al.*, 2025] He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. Instructmol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 354–379, 2025.

[Chang and Ye, 2024] Jinho Chang and Jong Chul Ye. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications*, 15(1):2323, 2024.

[Chiang *et al.*, 2023] Zhihao Chiang, Zhuohan Zhu, Xinyang Zeng, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[Christofidellis *et al.*, 2023] Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 6140–6157, 2023.

[Comanici *et al.*, 2025] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, et al. 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*, pages 4171–4186, 2019.

[Dey *et al.*, 2025] Vishal Dey, Xiao Hu, and Xia Ning. $\texttt{GeLLM}^{3O}$: Generalizing large language models for multi-property molecule optimization. *arXiv preprint arXiv:2502.13398*, 2025.

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

[Edwards *et al.*, 2021] Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607, 2021.

[Edwards *et al.*, 2022] Carl Edwards, Tuan Lai, et al. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, 2022.

[Elton *et al.*, 2019] Daniel C Elton, Zois Boukouvalas, Mark D Fuge, and Peter W Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, 4(4):828–849, 2019.

[Fang *et al.*, 2022] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2022.

[Fang *et al.*, 2023] Yin Fang, Xiaozhuan Liang, et al. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations*, 2023.

[Fu *et al.*, 2024] Li Fu, Shaohua Shi, Jiacai Yi, et al. Admetlab 3.0: an updated comprehensive online admet prediction platform enhanced with broader coverage, improved performance, api functionality and decision support. *Nucleic acids research*, 52(W1):W422–W431, 2024.

[GLM *et al.*, 2024] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[Grisoni, 2023] Francesca Grisoni. Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, 79:102527, 2023.

[Guo *et al.*, 2025] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[Hu *et al.*, 2019] Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[Irwin *et al.*, 2012] John J Irwin, Teague Sterling, Michael M Mysinger, Erin S Bolstad, and Ryan G Coleman. Zinc: a free tool to discover chemistry for

biology. *Journal of chemical information and modeling*, 52(7):1757–1768, 2012.

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

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

[Liu *et al.*, 2019] Shengchao Liu, Demirel, et al. N-gram graph: Simple unsupervised representation for graphs, with applications to molecules. *Advances in neural information processing systems*, 32, 2019.

[Liu *et al.*, 2023a] Zequn Liu, Wei Zhang, Yingce Xia, et al. MolXPT: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1606–1616, 2023.

[Liu *et al.*, 2023b] Zhiyuan Liu, Sihang Li, et al. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*, 2023.

[Liu *et al.*, 2024] Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in biology and medicine*, 171:108073, 2024.

[Loshchilov, 2017] I Loshchilov. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Ouyang *et al.*, 2024] Long Ouyang, Jeff Wu, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2024.

[Pei *et al.*, 2023] Qizhi Pei, Wei Zhang, et al. 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*, pages 1102–1123, 2023.

[Petroni *et al.*, 2019] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, et al. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[Rong *et al.*, 2020] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 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, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.

[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.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[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.

[Wu *et al.*, 2018] Zhenqin Wu, Bharath Ramsundar, Feinberg, et al. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[Yang and others, 2019] K. Yang et al. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019.

[Yang *et al.*, 2024] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

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

[Zhang *et al.*, 2023] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[Zhang *et al.*, 2024] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

[Zhao *et al.*, 2025] Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Yi Xia, Bo Chen, Hongshen Xu, Zichen Zhu, Su Zhu, et al. Developing chemdfm as a large language foundation model for chemistry. *Cell Reports Physical Science*, 6(4), 2025.

[Zhavoronkov, 2018] Alex Zhavoronkov. Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. *Molecular Pharmaceutics*, 15(10):4311–4313, 2018.

[Zheng *et al.*, 2024] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.