

# InstructMol: Multi-Modal Integration for Building a Versatile and Reliable Molecular Assistant in Drug Discovery

He Cao<sup>1,2\*</sup>, Zijing Liu<sup>1</sup>, Xingyu Lu<sup>1,3</sup>, Yuan Yao<sup>2</sup>, Yu Li<sup>1†</sup>

<sup>1</sup>International Digital Economy Academy (IDEA)

<sup>2</sup>Hong Kong University of Science and Technology

<sup>3</sup>Tsinghua Shenzhen International Graduate School, Tsinghua University

hcaoaf@connect.ust.hk luxy22@mails.tsinghua.edu.cn

yuany@ust.hk {liuzijing, liyu}@idea.edu.cn

## Abstract

The rapid evolution of artificial intelligence in drug discovery encounters challenges with generalization and extensive training, yet Large Language Models (LLMs) offer promise in reshaping interactions with complex molecular data. Our novel contribution, **InstructMol**<sup>1</sup>, a multi-modal LLM, effectively aligns molecular structures with natural language via an instruction-tuning approach, utilizing a two-stage training strategy that adeptly combines limited domain-specific data with molecular and textual information. **InstructMol** showcases substantial performance improvements in drug discovery-related molecular tasks, surpassing leading LLMs and significantly reducing the gap with specialists, thereby establishing a robust foundation for a versatile and dependable drug discovery assistant.

## 1 Introduction

The drug discovery process, from target identification to clinical trials, requires substantial investments in time and expertise for optimized exploration of chemical spaces (Coley, 2020). Artificial intelligence-driven drug discovery (AIDD) facilitates a data-driven modeling approach (Kim et al., 2021; Rifaioglu et al., 2018; Askr et al., 2022; Feng et al., 2024) and helps to understand the complex molecular space, reducing iterative testing and minimizing failure rates. Previous approaches involved employing task-specific models trained on labeled data, which had restricted adaptability and required laborious training for individual tasks. The advent of Large Language Models (LLMs (Devlin et al., 2019; Raffel et al., 2019; Brown et al., 2020)) like

ChatGPT (OpenAI, 2023a), trained through self-supervised learning on a large amount of unlabeled text data, has shown strong generalization capabilities across various tasks. Additionally, these models can attain professional-level proficiency in specific domains through proper fine-tuning. Hence, developing a ChatGPT-like molecular assistant AI can revolutionize human interactions with complex molecule structures. A unified model can address various needs, such as understanding molecule structures, answering drug-related queries, aiding synthesis planning, facilitating drug repurposing, etc., as shown in Figure 1.

Numerous studies have explored multimodal LLMs for visual understanding (Liu et al., 2023b; Ye et al., 2023; Zhu et al., 2023). However, when it comes to the domain of molecular research, there are several **challenges** that need to be addressed, including:

- Crafting a molecule representation integrates with LLMs alongside textual modalities;
- Requiring extensive datasets encompasses molecule structures, inherent properties, reactions, and annotations related to biological activities;
- Developing an effective training paradigm that guides LLMs in utilizing molecular representations and adapting to various tasks.

Several prior studies (Liang et al., 2023; Luo et al., 2023c; Fang et al., 2023) have fine-tuned generalist LLMs to develop foundational models within the molecular domain. Despite their enhancement to the original generalist LLM, these preceding works have unveiled several **issues**:

- Insufficient alignment between modalities.
- The consideration of an optimal molecular structure encoder remains unexplored.

\*Work done during an internship at IDEA.

† Corresponding authors: Yu Li (liyu@idea.edu.cn)

<sup>1</sup><https://github.com/IDEA-XL/InstructMol>

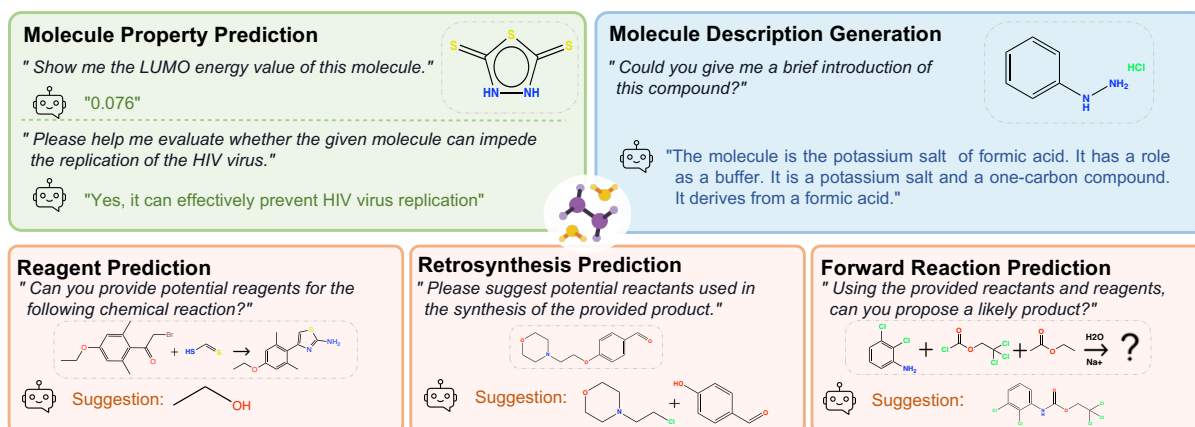


Figure 1: Empowering LLMs with molecular modalities to unlock the drug discovery domain and serve as assistants in molecular research.

- A rudimentary design of the training pipeline neglects the update of LLMs' knowledge.

These issues lead to a significant disparity in the performance of current AI assistants across various practical tasks compared to traditional specialist models.

To address these problems, we introduce **InstructMol** (Figure 2), a multi-modality instruction-tuning-based LLM. This model aligns molecular graphs and chemical sequential modalities with humans' natural language. Using a calibrated collection of molecule-related instruction datasets and a two-stage training scheme, **InstructMol** effectively leverages the pre-trained LLM and molecule graph encoder for molecule-text alignment. In the first alignment pretraining stage, we employ molecule-description pairs to train a lightweight and adaptable interface, which is designed to project the molecular node-level representation into the textual space that the LLM can understand. Subsequently, we finetune with multiple task-specific instructions. During this process, we freeze the molecule graph encoder and train low-rank adapters (LoRA (Hu et al., 2021)) on the LLM to adapt our model to various scenarios. This efficient approach enables the seamless integration of molecular and textual information, promoting the development of versatile and robust cognitive abilities in the molecular domain.

To illustrate the capabilities of our model, we perform experiments that span three facets of drug discovery-related tasks, including compound property prediction, molecule description generation, and analysis of chemical reactions involving compounds. These tasks serve as robust benchmarks to assess the model's ability to deliver useful and

accurate knowledge feedback in practical drug discovery scenarios. The results in all experiments consistently indicate that our model significantly improves the performance of LLMs in tasks related to the understanding and design of molecular compounds. Consequently, this advance effectively reduces the disparity with specialized models. Our main contributions can be summarized as follows:

- We introduce **InstructMol**, a molecular-related multi-modality LLM, representing a pioneering effort in bridging the gap between molecular and textual information.
- In the context of a scarcity of high-quality annotated data in the drug discovery domain, our approach strives to efficiently extract molecular representations (targets on **Issue2**). Employing a two-stage instruction tuning paradigm enhances the LLM's understanding of molecular structural and sequential knowledge (targets on **Issue1** and **Issue3**).
- InstructMol enables swift fine-tuning, generating lightweight checkpoints (used as plugins) for cross-modality tasks. It provides the flexibility to load or combine functionalities through plugins, retaining the open dialogue and reasoning capabilities of a general LLM.
- We evaluate our model through multiple practical assessments, demonstrating its substantial improvement compared to state-of-the-art LLMs. Our work lays the foundation for creating a versatile and reliable molecular research assistant in the drug discovery domain.

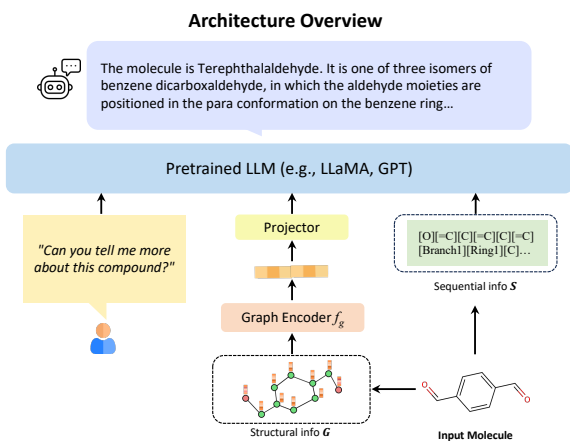


Figure 2: Overview of **InstructMol** model architecture design and two-stage training paradigm. The example molecule in the figure is *Terephthalaldehyde* (Sonmez et al., 2012) (CID 12173).

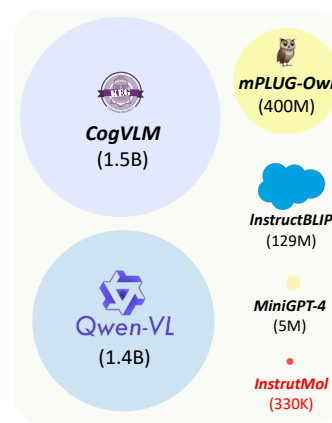
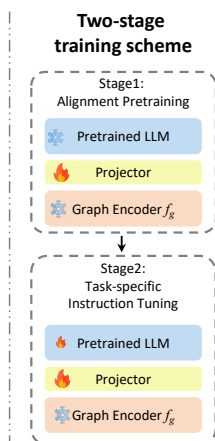


Figure 3: Comparison of biomolecule-domain molecule-text dataset scale with existing general domain vision-language datasets.

## 2 Related Work

### 2.1 Multimodal Instruction Tuning

There have been notable advancements in LLMs (OpenAI, 2023a; Touvron et al., 2023a,b; Chiang et al., 2023; Zeng et al., 2022a; Anil et al., 2023) achieved through scaling up model and data size. Consequently, LLMs have shown remarkable performances in zero/few-shot NLP tasks (OpenAI, 2023a; Wei et al., 2021; Ouyang et al., 2022). A key technique in LLMs is instruction tuning, where pre-trained LLMs are fine-tuned on instruction-formatted datasets (Wei et al., 2021), allowing them to generalize to new tasks. Recently, with the emergence of large foundation models in various domains, several efforts have been made to transition from unimodal LLMs to multimodal LLMs (MLLMs) (OpenAI, 2023b; Liu et al., 2023b; Zhu et al., 2023; Ye et al., 2023; Bai et al., 2023). The primary research on multimodal instruction tuning (M-IT) includes the following (Yin et al., 2023): *Constructing effective M-IT datasets* (adapting existing benchmarks datasets (Zhu et al., 2023; Liu et al., 2023b; Dai et al., 2023) or using self-instruction (Liu et al., 2023b; Wang et al., 2023; Li et al., 2023a; Zhang et al., 2023)), *Bridging diverse modalities* (project-based (Liu et al., 2023b; Li et al., 2023a; Pi et al., 2023) and query-based (Wang et al., 2023; Zhu et al., 2023; Ye et al., 2023)) and *Employing reliable evaluation methods* (GPT-scoring (Liu et al., 2023b; Li et al., 2023a; Chen et al., 2023; Luo et al., 2023a), manual scoring (Ye et al., 2023; Yang et al., 2023), or closed-set measurement (Liu et al., 2023b; Li et al., 2023a; Zhu et al., 2023; Luo et al., 2023a; Zhu et al., 2023;

Dai et al., 2023; Chen et al., 2023)). Most current MLLM research focuses on integrating vision and language while combining other modalities (e.g., graphs (Tang et al., 2023; Liu et al., 2023c)) with natural language remains nascent.

### 2.2 Molecule Foundation Models

The foundation models, trained on vast unlabeled data, serve as a paradigm for adaptable AI systems across diverse applications. In the single modality domain, researchers are exploring the molecule representations from diverse sources, such as 1D sequences (e.g., SMILES (Chithrananda et al., 2020; Irwin et al., 2021; Wang et al., 2019)), 2D molecular graphs (Wang et al., 2021; Hu et al., 2019; You et al., 2020), 3D geometric conformations (Stärk et al., 2021; Liu et al., 2021; Stärk et al., 2021), or textual information from biomedical literature (Gu et al., 2020; Lee et al., 2019; Beltagy et al., 2019). In the realm of multimodal analysis, research initiatives employ diverse approaches. These include encoder-decoder models to establish intermodal bridges (Edwards et al., 2022; Christofidellis et al., 2023; Lu and Zhang, 2022a), joint generative modeling of SMILES and textual data (Zeng et al., 2022b), and the adoption of contrastive learning for integrating molecular knowledge across varying modalities (Su et al., 2022; Luo et al., 2023b; Liu et al., 2022, 2023d).

### 2.3 Molecule-related LLMs

Given the rapid progress in LLMs, some researchers are considering developing ChatGPT-like AI systems for drug discovery. Their goal is to offer guidance for optimizing lead compounds, accu-

rately predicting drug interactions, and improving the comprehension of structure-activity relationships (Liang et al., 2023). Several initiatives have already commenced to create instruction datasets within the biomolecular domain (Fang et al., 2023; Lu et al., 2024). They aim to utilize instruction tuning techniques to enable LLMs, initially trained on general domain data, to acquire knowledge about biomolecular science (Wu et al., 2023; Luo et al., 2023c). Additionally, other researchers are investigating methods to align structural data with textual information, bridging the gap between biological data and natural language (Luo et al., 2023c; Liang et al., 2023; Cao et al., 2024b).

**Remark.** Our work involves molecule foundation models and multimodal language models (LLMs). It uses an efficient molecule graph encoder to capture structural information and integrates it with sequential data into a generalist LLM. **InstructMol** enables the LLM to understand molecule representations and generalize to various molecular tasks.

### 3 Method

#### 3.1 Multimodal Instruction Tuning

Instruction tuning refers to finetuning pretrained LLMs on instruction datasets, enabling generalization to specific tasks by adhering to new instructions. Multimodal instruction tuning integrates modalities like images and graphs into an LLM, expanding the model’s capability to accommodate multiple modalities.

A multimodal instruction tuning sample comprises an instruction  $I$  (e.g., "Describe the compound in detail") and an input-output pair. In the context of our study, the input is one or more modalities derived from a molecule (e.g., molecule graph and sequence), collectively denoted as  $M$ . The output  $R$  represents the textual response to the instruction conditioned on the input. The model aims to predict an answer given the instruction and multimodal input:  $\tilde{R} = f(I, M; \theta)$ , where  $\theta$  are the parameters of MLLM. The training objective is typically the same auto-regressive objective as the LLM pre-training stage, which can be expressed as:  $\mathcal{L}(\theta) = -\sum_{i=1}^L \log p(R_i|I, M, R_{<i}; \theta)$ , where  $L$  is the target  $R$ ’s token length.

#### 3.2 Construction of Molecular Instruction

**Data Collection.** In the field of biomolecular research, there is a noticeable scarcity of molecular datasets with comprehensive text annotations when

compared to the vision-language domain, as depicted in Figure 3. While it is possible to construct instruction datasets in general domains by adapting benchmarks or using self-instruction, the application of these methods in the biomolecular domain presents challenges. This difficulty arises from two main factors: 1) biomolecular domain annotation demands expert knowledge and entails substantial complexity; 2) the knowledge within this domain spans a broad range of subjects, including structural biology, computational chemistry, and chemical synthesis processes.

In our efforts, we have gathered recent open-access text-molecule pairs datasets and also independently constructed a portion of instruction data suitable for property prediction. Table 5 illustrates the composition of the data utilized during the two-stage training process.

**Molecule Input.** We utilize both the structure and sequence information of a molecule. We encode the structural information of a molecule as a graph, denoted by  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{X})$ , where  $\mathcal{V}$  is the set of atoms (nodes) and  $|\mathcal{V}| = N$  is the total number of atoms. The set of edges  $\mathcal{E}$  includes all chemical bonds, and  $\mathcal{A} \in \mathbb{R}^{N \times N}$  is the adjacency matrix. Additionally,  $\mathcal{X} \in \mathbb{R}^{N \times F}$  encompasses attributes associated with each node, where  $F$  is the feature dimension. With a Graph Encoder  $f_g$ , we extract a graph representation  $\mathbf{Z}_G \in \mathbb{R}^{N \times d}$  at the node level, effectively describing the inherent structure of the molecule. Simultaneously, we consider encoding the sequential information of the molecule, denoted as  $S$ , as a supplementary source of structural information. To enhance the robustness of sequential molecular descriptors and mitigate syntactic and semantic invalidity present in SMILES (Weininger, 1988), we employ SELFIES (Krenn et al., 2019) as  $S$ , which is designed for mapping each token to a distinct structure or reference.

**Input Formulation.** We formulate a molecule-text pair ( $\mathbf{X}_M$  &  $\mathbf{X}_c$ ) to the corresponding instruction-following version like Human:  $\mathbf{X}_I <\text{mol}> \mathbf{X}_M <\text{STOP}>$  Assistant:  $\mathbf{X}_A <\text{STOP}>$ . The  $\mathbf{X}_M$  represents the molecule, including the molecule graph  $\mathbf{X}_G$  and optionally the SELFIES  $\mathbf{X}_S$ .  $\mathbf{X}_I$  denotes for the instruction and  $\mathbf{X}_A$  is the answer. For a given answer sequence of length  $L$ , our optimization objective is to maximize the probability of

generating the target answers  $X_A$  by maximizing:

$$p(\mathbf{X}_A|\mathbf{X}_M, \mathbf{X}_I) = \prod_{i=1}^L p_{\theta}(x_i | \mathbf{X}_G \parallel \mathbf{X}_S, \mathbf{X}_I, \mathbf{X}_{A,<i}). \quad (1)$$

To diversify  $\mathbf{X}_I$ , we craft clear task descriptions and use GPT-3.5-turbo to generate varied questions, enhancing instructions’ robustness. Note that we simply concatenate  $\mathbf{X}_G$  and  $\mathbf{X}_S$  along the length-dimension. More complex fusion methods require additional loss designs for supervision (Liu et al., 2023d; Luo et al., 2023b), but here we prioritize simplicity.

### 3.3 Architecture

**Molecular Encoder.** The molecular graph encoder,  $f_g$ , needs to efficiently extract node representations while preserving the molecular graph’s connectivity information. It is crucial that  $f_g$  inherently establishes a pre-alignment in the representation space with the text space to facilitate  $\mathbf{Z}_G$  in the following alignment stage. Taking inspiration from common practices in the Vision Large Language Models (VLLM) domain (Bai et al., 2023; Liu et al., 2023b; Ye et al., 2023), where models like ViT initialized from CLIP (Radford et al., 2021) serve as vision encoders, we optimize for MoleculeSTM’s graph encoder as  $f_g$  (Liu et al., 2022), instead of GraphMVP used by prior methodologies (Liang et al., 2023; Luo et al., 2023c). The MoleculeSTM graph encoder model is obtained through molecular-textual contrastive training, mitigating the requirement for an extensive amount of paired data during training to align different modalities.

**Light-weight Alignment Projector.** To map graph features into the word embedding space, we utilize a trainable projection matrix  $\mathbf{W}$  to transform  $\mathbf{Z}_G$  into  $\mathbf{X}_G$ , ensuring that it has the same dimension as the word embedding space. Since the selected  $f_g$  has undergone partial alignment with the text through contrastive training, we believe a straightforward linear projection will meet the subsequent alignment needs. For approaches like gated cross-attention (Alayrac et al., 2022), Q-former (et.al., 2023), or position-aware vision-language adapters (Bai et al., 2023), they require a large number of pairs for pretraining alignment, which is typically unavailable in the biomolecular domain. We therefore do not explore these more complex alignment methods.

**Large Language Model.** InstructMol incorporates a pre-trained LLM as its foundational component.

We optimize for Vicuna-7B (Chiang et al., 2023) as the initialized weights, which is derived from LLaMA (Touvron et al., 2023a) through supervised instruction finetuning.

### 3.4 Two-Stage of Instruction Tuning

As illustrated in Figure 2, the training process of InstructMol consists of two stages: alignment pre-training and instruction fine-tuning training.

**Alignment Pretraining.** In the first stage, we aim to align the modality of molecules with text, ensuring that the LLMs can perceive both the structural and sequential information of molecules and integrate molecular knowledge into their internal capabilities.

We primarily employ a dataset consisting of molecule-text pairs sourced from PubChem (Kim et al., 2022). Each molecule structure is associated with a textual description elucidating chemical and physical properties or high-level bioactivity information. The construction of the PubChemDataset predominantly follows the MoleculeSTM (Liu et al., 2022) pipeline. We meticulously remove molecules with invalid descriptions and syntactic errors in their molecular descriptors. To ensure fairness, we also eliminate compounds that might appear in the downstream molecule-caption test set. This results in a dataset of 330K molecule-text pairs. Subsequently, we adopt a self-instruction-like approach to generate a diverse set of task descriptions as instructions.

During training, to prevent overfitting and leverage pre-trained knowledge, we freeze both the graph encoder and LLM, focusing solely on fine-tuning the alignment projector. After a few epochs of training, our aim is that the projector has successfully learned to map graph representations to graph tokens, aligning effectively with text tokens.

**Task-specific Instruction Tuning.** In the second stage, we target three distinct downstream scenarios. We advocate for task-specific instruction tuning to address the particular constraints inherent in various drug-discovery-related tasks. For *compound property prediction*, we utilize the quantum mechanics properties instruction dataset from Fang et al. (2023) for regression prediction and the MoleculeNet dataset (Wu et al., 2017) for property classification. For *chemical reaction analysis*, we incorporate forward reaction prediction, retrosynthesis analysis, and reagent prediction tasks, all derived from Fang et al. (2023). To assess the

model’s proficiency in translating between natural language and molecular expression, we integrate ChEBI-20 (Edwards et al., 2021) for the *molecule description generation task*. For each task, corresponding instruction templates are designed.

During training, we utilize the checkpoint of the alignment projector that was trained in the first stage as initialization. We only keep the molecular encoder  $f_g$  frozen and continue to update the pre-trained weights of the projector and the LLM. To adapt the LLM effectively for diverse tasks, we employ low-rank adaptation (i.e., LoRA (Hu et al., 2021)), opting against full-tuning to mitigate potential forgetting issues. In practical applications, we have the flexibility to substitute different adaptors based on specific scenario requirements or combine multiple adaptors to integrate knowledge, thereby showcasing the model’s modularization capabilities. Moreover, LoRA allows the LLM to retain the inherent capacity for common-sense reasoning in dialogue (as shown in Table 13).

## 4 Experiments

We use a graph neural network as the molecule graph encoder ( $f_g$ ) which is initialized with the MoleculeSTM graph encoder, pre-trained through molecular graph-text contrastive learning. We employ Vicuna-v-1.3-7B (Chiang et al., 2023) as the base LLM. More specifically, **InstructMol+GS** denotes we inject both molecular graph tokens and sequence tokens into the input, while **InstructMol+G** means only incorporates graph tokens. For Instruct-S, which utilizes only a 1D molecular sequence as input, it corresponds to the fine-tuning of the base large language model, Vicuna-7B, directly on downstream tasks. In the following sections, the results of Vicuna-v1.3-7B will consistently be used to represent the performance of Instruct-S. Implementation details about model settings and training hyper-parameters can be referred to Appendix B.

### 4.1 Property Prediction Task

**Experiment Setup.** Property prediction intends to forecast a molecule’s intrinsic physical and chemical properties from its structural or sequential characteristics. In the context of the regression task, we undertake experiments on the Property Prediction dataset from Fang et al. (2023), where the objective is to predict the quantum mechanic’s properties of a given molecule, specifically including HOMO, LUMO, and the HOMO-LUMO gap (Ramakrish-

nan et al., 2014b). For the classification task, we incorporate three binary classification datasets of molecular biological activity, namely BACE, BBBP, and HIV. In classification, all dataset samples are converted into an instruction format and we use the recommended splits from (Ramsundar et al., 2019). Each item comprises an instruction explaining the property for prediction and the representation of the molecule. Subsequently, models are tasked with generating a single prediction (“yes” or “no”). Scaffold splits are used for the classification task, and the experiments are conducted with three random seeds, yielding low variances in the reported mean values.

METHOD	HOMO ↓	LUMO ↓	$\Delta\epsilon$ ↓	AVG ↓
<i>LLM Based Generalist Models</i>				
Alpaca <sup>†</sup> (Taori et al., 2023)	-	-	-	322.109
Baize <sup>†</sup> (Xu et al., 2023)	-	-	-	261.343
Galactica <sup>†</sup> (Taylor et al., 2022)	-	-	-	0.568
LLama-2-7B (5-shot ICL)	0.7367	0.8641	0.5152	0.7510
Vicuna-13B (5-shot ICL)	0.7135	3.6807	1.5407	1.9783
Mol-Instruction	0.0210	0.0210	0.0203	0.0210
<b>InstructMol-G</b>	0.0060	0.0070	0.0082	0.0070
<b>InstructMol-GS</b>	<b>0.0048</b>	<b>0.0050</b>	<b>0.0061</b>	<b>0.0050</b>

Table 1: Results (MAE in hartree unit) for QM9 property regression tasks. <sup>†</sup> denotes few-shot in-context learning (ICL) results from Fang et al. (2023).  $\Delta\epsilon$  denotes HOMO-LUMO energy gap.

METHOD	BACE ↑	BBBP ↑	HIV ↑
# MOLECULES	1513	2039	41127
<i>Specialist Models</i>			
ChemBERTa v2 (Walid et al., 2022)	73.5	69.8	79.3
DMP(TF+GNN) (Jinhua et al., 2023)	89.4	77.8	81.4
KV-PLM (Zeng et al., 2022b)	78.5	70.5	71.8
GraphCL (You et al., 2020)	75.3	69.7	78.5
GraphMVP-C (Liu et al., 2021)	81.2	72.4	77.0
MoMu (Su et al., 2022)	76.7	70.5	75.9
MolFM (Luo et al., 2023b)	83.9	72.9	78.8
Uni-Mol (Zhou et al., 2023)	85.7	72.9	80.8
<i>LLM Based Generalist Models</i>			
Galactica-6.7B	58.4	53.5	72.2
Vicuna-v1.5-13b-16k (4-shot)	49.2	52.7	50.5
Vicuna-v1.3-7B*	68.3	60.1	58.1
LLama-2-7B-chat*	74.8	65.6	62.3
MolCA(1D) (Liu et al., 2023f)	79.3	70.8	-
MolCA(1D + 2D) (Liu et al., 2023f)	79.8	70.0	-
<b>Instruct-G</b>	<b>84.3</b> ( $\pm 0.6$ )	<b>68.6</b> ( $\pm 0.3$ )	<b>74.0</b> ( $\pm 0.1$ )
<b>Instruct-GS</b>	<b>82.1</b> ( $\pm 0.1$ )	<b>72.4</b> ( $\pm 0.3$ )	<b>68.9</b> ( $\pm 0.3$ )

Table 2: ROC-AUC of molecular property prediction tasks (classification) on MoleculeNet (Wu et al., 2017) benchmarks. \*: use LoRA tuning. We indicate the best performance among domain specialist models by underlining the results, while the best performance among LLM-based generalist models is highlighted in bold.

**Results.** Our models are compared against baselines on the test set for regression, measured by Mean Absolute Error (MAE) in Table 1. Compared to previous single-modal instruction-tuned LLM-based methods (Fang et al., 2023), InstructMol demonstrates a further improvement in the regression task. ROC-AUC scores for classifica-

MODEL	BLEU-2 $\uparrow$	BLEU-4 $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-2 $\uparrow$	ROUGE-L $\uparrow$	METEOR $\uparrow$
<i>Specialist Models</i>						
MolT5-base (Edwards et al., 2022)	0.540	0.457	0.634	0.485	0.568	0.569
MoMu (MolT5-base) (Su et al., 2022)	0.549	0.462	-	-	-	0.576
MolFM (MolT5-base) (Luo et al., 2023b)	0.585	0.498	0.653	0.508	0.594	0.607
MolXPT (Liu et al., 2023e)	0.594	0.505	0.660	0.511	0.597	0.626
GIT-Mol-graph (Liu et al., 2023d)	0.290	0.210	0.540	0.445	0.512	0.491
GIT-Mol-SMILES (Liu et al., 2023d)	0.264	0.176	0.477	0.374	0.451	0.430
GIT-Mol-(graph+SMILES) (Liu et al., 2023d)	0.352	0.263	0.575	0.485	0.560	0.430
MolCA, Galac <sub>1.3B</sub> (Liu et al., 2023f)	0.620	0.531	0.681	0.537	0.618	0.651
Text+Chem T5-augm-base (Christofidellis et al., 2023)	0.625	0.542	0.682	0.543	0.622	0.648
<i>Retrieval Based LLMs</i>						
GPT-3.5-turbo (10-shot MolReGPT) (Li et al., 2023b)	0.565	0.482	0.623	0.450	0.543	0.585
GPT-4-0314 (10-shot MolReGPT) (Li et al., 2023b)	0.607	0.525	0.634	0.476	0.562	0.610
<i>LLM Based Generalist Models</i>						
GPT-3.5-turbo (zero-shot) (Li et al., 2023b)	0.103	0.050	0.261	0.088	0.204	0.161
BioMedGPT-10B (Luo et al., 2023c)	0.234	0.141	0.386	0.206	0.332	0.308
Mol-Instruction (Fang et al., 2023)	0.249	0.171	0.331	0.203	0.289	0.271
<b>InstructMol-G</b>	0.466	0.365	0.547	0.365	0.479	0.491
<b>InstructMol-GS</b>	<b>0.475</b>	<b>0.371</b>	<b>0.566</b>	<b>0.394</b>	<b>0.502</b>	<b>0.509</b>

Table 3: Results of molecular description generation task on the test split of ChEBI-20.

tion outcomes are presented in Table 2. In comparison to LLM-based generalist models, both the Galactica (Taylor et al., 2022) series models trained on an extensive scientific literature dataset and the single-modality LLM fine-tuned with task-specific instructions (Fang et al., 2023), InstructMol demonstrates consistent improvements in accuracy across the three task datasets. However, our predictive results still exhibit some disparity compared to expert models (Zhou et al., 2023; Liu et al., 2021) specifically trained on a vast molecule structure dataset. Further, InstructMol performs worse than GIN on the imbalanced HIV dataset with a long-tail distribution. Previous research (Kandpal et al., 2023) highlights LLMs’ challenges in learning long-tail knowledge. To tackle this, strategies like resampling or class reweighting can be employed.

## 4.2 Molecule Description Generation Task

**Experiment Setup.** Molecule description generation encapsulates a comprehensive molecule depiction, covering its structure, properties, biological activity, and applications based on molecular descriptors. This task is more complex than classification or regression, providing a robust measure of the model’s understanding of molecules. We convert the training subset of the ChEBI-20 dataset (Edwards et al., 2021) into an instructional format and subsequently perform fine-tuning based on these instructions. Our assessment uses evaluation metrics aligned with (Edwards et al., 2022).

**Baselines.** Three kinds of models are used as baselines, including: 1) MolT5-like expert models (Edwards et al., 2022; Liu et al., 2023e) and the models employing MolT5 as a decoder (Su et al., 2022;

Luo et al., 2023b; Liu et al., 2023d; Christofidellis et al., 2023), 2) models based on retrieval methods that utilize ChatGPT/GPT-4 as a foundational component (Li et al., 2023b), 3) other models derived through instruction-tuning with LLMs to achieve generalist unimodal (Fang et al., 2023) and multimodalities (Luo et al., 2023c) capabilities.

**Results.** Table 3 presents the overall results for molecule description generation. Our model outperforms other generalist LLM-based models in generating precise, contextually relevant molecule descriptions. We observe that incorporating both molecule structural information and sequential information in the input yields higher-quality results ( $\sim 2\%$  improvement) than providing structural information alone. While expert models demonstrate better efficacy in comparison, it is noteworthy that they are constrained by their training schemes and lack the versatile capabilities inherent in our approach. Retrieval methods, supported by ChatGPT/GPT-4, demonstrate strong capabilities. Our future efforts will focus on integrating these methods to improve the accuracy and credibility of generated content.

## 4.3 Chemical Reaction-related task

**Experiment Setup.** Traditionally, identifying chemical reactions relied on intuition and expertise. Integrating deep learning for predicting reactions can accelerate research and improve drug discovery. The general format of a chemical reaction is "reactant  $\rightarrow$  reagent  $\rightarrow$  product". Here we mainly focus on three tasks: 1) *Forward Reaction Prediction*: predict the probable product(s) given specific reactants and reagents; 2) *Reagent Predic-*

*tion*: ascertain the suitable catalysts, solvents, or ancillary substances required for a specific chemical reaction given reactant(s) and product(s); 3) *Retrosynthesis*: anticipate deducing potential precursor molecule(s) from given product(s).

We utilize the dataset sourced from Fang et al. (2023), training it on the pre-defined training split and evaluating its performance on the test set. The performance is assessed by metrics like Fingerprint Tanimoto Similarity (FTS), BLEU, Exact Match, and Levenshtein distance to measure the similarity between ground truth and prediction. We also measure the validity of predicted molecules using RDKit.

**Results.** Table 4 reports the outcomes of tasks related to chemical reactions. It is evident that **InstructMol** outperforms the baselines significantly. The results obtained by generalist LLMs are derived from Fang et al. (2023), and they exhibit a pronounced inability to comprehend any chemical reaction prediction task, struggling to generate valid molecule(s) as answers. Mol-Instruction (Fang et al., 2023), employing Llama2 (Touvron et al., 2023b) as the base LLM, is jointly trained on multiple molecule-oriented instruction datasets. In addition, we supplement this by adopting the same training settings but exclusively training on chemical reaction-related datasets. Through comparison, InstructMol, as a multi-modality LLM, demonstrates a superior understanding of the task compared to single-modality models, confirming its effectiveness as a chemical reaction assistant.

#### 4.4 Ablation Studies

In this subsection, we conduct an ablation study to investigate the architecture and training scheme design of our proposed framework. We explore variations from several perspectives and validate them on the task of molecule description generation. The ablation results are presented in Appendix Table 10 as follows: **1) Employing an MLP connector instead of a linear projector.** Drawing inspiration from the observations made in (Liu et al., 2023a), we attempt to change the alignment projector to a two-layer MLP, demonstrating an enhancement in the model’s multimodal capabilities. **2) Scaling up the LLM to 13B.** The results indicate that scaling up the LLM only yields minor improvements. Thus, it substantiates the assertion that, for specific domains characterized by dataset scarcity, employing a 7B size model is sufficiently efficient for modeling. **3) Replacing the**

**graph encoder  $f_g$  with a single-modality module** (i.e., GraphMVP (Liu et al., 2021) with the same parameter size and architecture as we used). The results affirm our perspective: utilizing an encoder pre-aligned with text enhances the effectiveness of modality alignment. **4) Skipping alignment stage-1.** We included a comparison where stage-1 was skipped entirely. The results demonstrate that separating projector training (stage-1) from downstream fine-tuning (stage-2) yields better performance. **5) Freezing the LLM in the second stage.** Adopting a strategy akin to BioMedGPT10B (Luo et al., 2023c) and DrugChat (Liang et al., 2023), we choose not to update LLM weights in the second stage. The training outcomes reveal challenges in convergence and an inability to complete normal inference, thus demonstrating the necessity for the instruct-tuning stage to adapt LLM knowledge to the specific task.

## 5 Discussion and Conclusion

**Conclusion.** We propose InstructMol, a novel multi-modality foundational model that connects molecular modalities with human natural language. By integrating structural and sequential information of molecules into LLMs through a dual-stage alignment pre-training and instruction tuning paradigm, we enhance the general LLM’s capacity to comprehend and interpret molecular information, specifically in drug discovery tasks. Extensive experimental evaluation confirms the effectiveness of our model architecture and training approach, demonstrating its potential for practical applications in the field of drug discovery.

**Future Work.** Integrating multiple modalities with LLMs significantly enhances molecular research within this domain and is a valuable direction to explore. However, several challenges exist. The scale and quality of relevant datasets are as good as those in the vision and language community. The lack of well-defined task objectives poses a challenge. A more scientifically robust evaluation is needed to address issues such as hallucinations in generation outputs.

## 6 Limitations

In our investigation, several limitations have emerged. Firstly, the scale and quality of the dataset pose significant constraints; the scarcity of high-quality annotated domain data may hinder the model’s ability to generalize across the diverse



MODEL	EXACT $\uparrow$	BLEU $\uparrow$	LEVENSHTEIN $\downarrow$	RDk FTS $\uparrow$	MACCS FTS $\uparrow$	MORGAN FTS $\uparrow$	VALIDITY $\uparrow$
<i>Reagent Prediction</i>							
Alpaca $\dagger$ (Taori et al., 2023)	0.000	0.026	29.037	0.029	0.016	0.001	0.186
Baize $\dagger$ (Xu et al., 2023)	0.000	0.051	30.628	0.022	0.018	0.004	0.099
ChatGLM $\dagger$ (Zeng et al., 2022a)	0.000	0.019	29.169	0.017	0.006	0.002	0.074
LLama $\dagger$ (Touvron et al., 2023a)	0.000	0.003	28.040	0.037	0.001	0.001	0.001
Vicuna $\dagger$ (Chiang et al., 2023)	0.000	0.010	27.948	0.038	0.002	0.001	0.007
Mol-Instruction (Fang et al., 2023)	0.044	0.224	23.167	0.237	0.364	0.213	1.000
LLama-7b* (Touvron et al., 2023a)(LoRA)	0.000	0.283	53.510	0.136	0.294	0.106	1.000
<b>InstructMol-G</b>	0.070	<b>0.890</b>	24.732	<b>0.469</b>	<b>0.691</b>	<b>0.426</b>	1.000
<b>InstructMol-GS</b>	<b>0.129</b>	0.610	<b>19.664</b>	0.444	0.539	0.400	1.000
<i>Forward Reaction Prediction</i>							
Alpaca $\dagger$ (Taori et al., 2023)	0.000	0.065	41.989	0.004	0.024	0.008	0.138
Baize $\dagger$ (Xu et al., 2023)	0.000	0.044	41.500	0.004	0.025	0.009	0.097
ChatGLM $\dagger$ (Zeng et al., 2022a)	0.000	0.183	40.008	0.050	0.100	0.044	0.108
LLama $\dagger$ (Touvron et al., 2023a)	0.000	0.020	42.002	0.001	0.002	0.001	0.039
Vicuna $\dagger$ (Chiang et al., 2023)	0.000	0.057	41.690	0.007	0.016	0.006	0.059
Mol-Instruction (Fang et al., 2023)	0.045	0.654	27.262	0.313	0.509	0.262	1.000
LLama-7b* (Touvron et al., 2023a)(LoRA)	0.012	0.804	29.947	0.499	0.649	0.407	1.000
Text+ChemT5 (Christofidellis et al., 2023)	0.454	0.602	26.545	0.729	0.773	0.700	0.851
MolecularTransformer (Schwaller et al., 2018)	0.0	0.476	45.979	0.761	0.673	0.540	1.000
<b>InstructMo-G</b>	0.153	0.906	20.155	0.519	0.717	0.457	1.000
<b>InstructMol-GS</b>	<b>0.536</b>	<b>0.967</b>	<b>10.851</b>	<b>0.776</b>	<b>0.878</b>	<b>0.741</b>	1.000
<i>Retrosynthesis</i>							
Alpaca $\dagger$ (Taori et al., 2023)	0.000	0.063	46.915	0.005	0.023	0.007	0.160
Baize $\dagger$ (Xu et al., 2023)	0.000	0.095	44.714	0.025	0.050	0.023	0.112
ChatGLM $\dagger$ (Zeng et al., 2022a)	0.000	0.117	48.365	0.056	0.075	0.043	0.046
LLama $\dagger$ (Touvron et al., 2023a)	0.000	0.036	46.844	0.018	0.029	0.017	0.010
Vicuna $\dagger$ (Chiang et al., 2023)	0.000	0.057	46.877	0.025	0.030	0.021	0.017
Mol-Instruction (Fang et al., 2023)	0.009	0.705	31.227	0.283	0.487	0.230	1.000
LLama-7b* (Touvron et al., 2023a)(LoRA)	0.000	0.283	53.510	0.136	0.294	0.106	1.000
Text+ChemT5 (Christofidellis et al., 2023)	0.033	0.314	88.672	0.457	0.469	0.350	0.632
Retroformer-untyped (Yao et al., 2022)	<b>0.536</b>	0.881	<b>10.277</b>	<b>0.865</b>	<b>0.904</b>	<b>0.830</b>	0.995
<b>InstructMol-G</b>	0.114	0.586	21.271	0.422	0.523	0.285	1.000
<b>InstructMol-GS</b>	0.407	<b>0.941</b>	13.967	0.753	0.852	0.714	1.000

Table 4: Results of chemical reaction tasks.  $\dagger$ : few-shot ICL results from (Fang et al., 2023). \*: use task-specific instruction data to fine-tune. Model indicates a domain expert method.

and intricate molecular landscapes encountered in real-world applications. Secondly, the integration and evaluation of multiple modalities have also revealed areas needing improvement. Further refinement is necessary to ensure robust alignment and utilization of different molecule modalities within the model, enhancing its capacity to interpret and generate responses accurately across the molecular domain. Lastly, our base LLM originates from a general-domain model. However, the absence of specialized LLMs tailored specifically for chemistry and molecular science, like models such as LLaMA, highlights the need for larger, more versatile domain-specific LLMs to enhance performance and expand applications. Addressing these challenges is pivotal for enhancing the model’s reliability and extending its utility in advancing drug discovery methodologies.

## 7 Potential Risks

The application of AI in drug discovery entails several potential risks. A primary concern is the potential misuse of AI to develop hazardous or illicit substances, which presents significant safety and

ethical challenges. Moreover, inaccuracies in AI-generated outputs could lead to hazardous chemical reactions if not thoroughly verified, posing risks of harm or damage to equipment. Dependence on AI-generated content heightens the risk of accidents and unsafe practices. Therefore, stringent oversight and rigorous adherence to ethical guidelines are essential to mitigate these risks and ensure the safe and responsible application of AI in drug discovery. Further insights into these issues and potential safeguard approaches can be found in recent literature (Wong et al., 2024; Cao et al., 2024a; Wang et al., 2024).

## Acknowledgements

This project was supported in part by Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, under Grant No. HTHZQSW-S-KCCYB-2023052, National Natural Science Foundation of China / Research Grants Council Joint Research Scheme Grant N\_HKUST635/20, and HKRGC Grant 16308321.

## References

- PubChem Structure Search. [https://pubchem.ncbi.nlm.nih.gov/search/help\\_search.html](https://pubchem.ncbi.nlm.nih.gov/search/help_search.html).
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). *ArXiv*, abs/2204.14198.
- Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D’iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Oleksandr Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. [Palm 2 technical report](#). *ArXiv*, abs/2305.10403.
- Heba Askr, Enas Elgeldawi, Heba Aboul Ella, Yaseen A.M.M. Elshaiar, Mamdouh M. Gomaa, and Aboul Ella Hassanien. 2022. [Deep learning in drug discovery: an integrative review and future challenges](#). *Artificial Intelligence Review*, 56:5975 – 6037.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A frontier large vision-language model with versatile abilities](#). *ArXiv*, abs/2308.12966.
- Satanjeev Banerjee and Alon Lavie. 2005. [Meteor: An automatic metric for mt evaluation with improved correlation with human judgments](#). In *IEEvaluation@ACL*.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Andrés M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2023. [Chemcrow: Augmenting large-language models with chemistry tools](#).
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher 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](#). *ArXiv*, abs/2005.14165.

- He Cao, Weidi Luo, Yu Wang, Zijing Liu, Bing Feng, Yuan Yao, and Yu Li. 2024a. Guide for defense (g4d): Dynamic guidance for robust and balanced defense in large language models. *arXiv preprint arXiv:2410.17922*.
- He Cao, Yanjun Shao, Zhiyuan Liu, Zijing Liu, Xiangru Tang, Yuan Yao, and Yu Li. 2024b. **PRESTO: Progressive pretraining enhances synthetic chemistry outcomes**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10197–10224, Miami, Florida, USA. Association for Computational Linguistics.
- Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. 2023. **X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages**. *ArXiv*, abs/2305.04160.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. **Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality**.
- Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. **Chemberta: Large-scale self-supervised pretraining for molecular property prediction**. *ArXiv*, abs/2010.09885.
- 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*.
- Connor W. Coley. 2020. **Defining and exploring chemical spaces**. *Trends in Chemistry*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. 2023. **Instructblip: Towards general-purpose vision-language models with instruction tuning**. *ArXiv*, abs/2305.06500.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. **Bert: Pre-training of deep bidirectional transformers for language understanding**. In *North American Chapter of the Association for Computational Linguistics*.
- Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. **Reoptimization of mdl keys for use in drug discovery**. *Journal of chemical information and computer sciences*, 42 6:1273–80.
- Carl N. Edwards, T. Lai, Kevin Ros, Garrett Honke, and Heng Ji. 2022. **Translation between molecules and natural language**. *ArXiv*, abs/2204.11817.
- Carl N. Edwards, Chengxiang Zhai, and Heng Ji. 2021. **Text2mol: Cross-modal molecule retrieval with natural language queries**. In *Conference on Empirical Methods in Natural Language Processing*.
- Li et.al. 2023. **Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models**. In *ICML*.
- Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. **Mol-instructions: A large-scale biomolecular instruction dataset for large language models**. *ArXiv*, abs/2306.08018.
- Bin Feng, Zequn Liu, Nanlan Huang, Zhiping Xiao, Haomiao Zhang, Srбуhi Mirzoyan, Hanwen Xu, Jiaran Hao, Yinghui Xu, Ming Zhang, et al. 2024. **A bioactivity foundation model using pairwise meta-learning**. *Nature Machine Intelligence*, 6(8):962–974.
- Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. **Domain-specific language model pretraining for biomedical natural language processing**. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Xiaodong Song, and Jacob Steinhardt. 2020. **Measuring massive multitask language understanding**. *ArXiv*, abs/2009.03300.
- J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. **Lora: Low-rank adaptation of large language models**. *ArXiv*, abs/2106.09685.

- Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. 2019. [Strategies for pre-training graph neural networks](#). *arXiv: Learning*.
- Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. 2021. [Chemformer: a pre-trained transformer for computational chemistry](#). *Machine Learning: Science and Technology*, 3.
- Zhu Jinhua, Xia Yingce, Wu Lijun, Xie Shufang, Zhou Wengang, Qin Tao, Li Houqiang, and Liu Tie-Yan. 2023. Dual-view molecular pre-training. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3615–3627.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.
- Jin Kim, Sera Park, Dongbo Min, and Wankyu Kim. 2021. [Comprehensive survey of recent drug discovery using deep learning](#). *International Journal of Molecular Sciences*, 22.
- Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Y. Zaslavsky, Jian Zhang, and Evan E. Bolton. 2022. [Pubchem 2023 update](#). *Nucleic acids research*.
- Sunghwan Kim, Paul A. Thiessen, Tiejun Cheng, Jian Zhang, Asta Gindulyte, and Evan E. Bolton. 2019. [Pug-view: programmatic access to chemical annotations integrated in pubchem](#). *Journal of Cheminformatics*, 11.
- Mario Krenn, Florian Hase, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2019. [Self-referencing embedded strings \(selfies\): A 100% robust molecular string representation](#). *Machine Learning: Science and Technology*, 1.
- Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. [Biobert: a pre-trained biomedical language representation model for biomedical text mining](#). *Bioinformatics*, 36:1234–1240.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. [Llava-med: Training a large language-and-vision assistant for biomedicine in one day](#). *ArXiv*, abs/2306.00890.
- Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao Wei, Hui Liu, Jiliang Tang, and Qing Li. 2023b. [Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective](#). *ArXiv*, abs/2306.06615.
- Youwei Liang, Ruiyi Zhang, Li Zhang, and Peng Xie. 2023. [Drugchat: Towards enabling chatgpt-like capabilities on drug molecule graphs](#). *ArXiv*, abs/2309.03907.
- Chin-Yew Lin. 2004. [Rouge: A package for automatic evaluation of summaries](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. [Improved baselines with visual instruction tuning](#). *ArXiv*, abs/2310.03744.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. [Visual instruction tuning](#). *ArXiv*, abs/2304.08485.
- Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen, Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang, Lichao Sun, Philip S. Yu, and Chuan Shi. 2023c. [Towards graph foundation models: A survey and beyond](#). *ArXiv*, abs/2310.11829.
- Peng Liu, Yiming Ren, and Zhixiang Ren. 2023d. [Git-mol: A multi-modal large language model for molecular science with graph, image, and text](#). *ArXiv*, abs/2308.06911.
- Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. [Multi-modal molecule structure-text model for text-based retrieval and editing](#). *ArXiv*, abs/2212.10789.
- Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. [Pre-training molecular graph representation with 3d geometry](#). *ArXiv*, abs/2110.07728.

- Zequan Liu, W. Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Yang Zhang, and Tiejian Liu. 2023e. [Molxpt: Wrapping molecules with text for generative pre-training](#). *ArXiv*, abs/2305.10688.
- Zhiyuan Liu, Sihang Li, Yancheng Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023f. [Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Jieyu Lu and Yingkai Zhang. 2022a. [Unified deep learning model for multitask reaction predictions with explanation](#). *Journal of chemical information and modeling*.
- Jieyu Lu and Yingkai Zhang. 2022b. [Unified deep learning model for multitask reaction predictions with explanation](#). *Journal of chemical information and modeling*.
- Xingyu Lu, He Cao, Zijing Liu, Shengyuan Bai, Leqing Chen, Yuan Yao, Hai-Tao Zheng, and Yu Li. 2024. [MoleculeQA: A dataset to evaluate factual accuracy in molecular comprehension](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3769–3789, Miami, Florida, USA. Association for Computational Linguistics.
- Gen Luo, Yiyi Zhou, Tianhe Ren, Shen Chen, Xiaoshuai Sun, and Rongrong Ji. 2023a. [Cheap and quick: Efficient vision-language instruction tuning for large language models](#). *ArXiv*, abs/2305.15023.
- Yi Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023b. [Molfm: A multi-modal molecular foundation model](#). *ArXiv*, abs/2307.09484.
- Yi Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023c. [Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine](#). *ArXiv*, abs/2308.09442.
- OpenAI. 2023a. ["chatgpt: A language model for conversational ai](#).
- OpenAI. 2023b. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. [Training language models to follow instructions with human feedback](#). *ArXiv*, abs/2203.02155.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.
- Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. [Detgpt: Detect what you need via reasoning](#). *ArXiv*, abs/2305.14167.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). In *International Conference on Machine Learning*.
- Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *J. Mach. Learn. Res.*, 21:140:1–140:67.
- Raghunathan Ramakrishnan, Pavlo O. Dral, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. 2014a. [Quantum chemistry structures and properties of 134 kilo molecules](#). *Scientific Data*, 1.
- Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. 2014b. [Quantum chemistry structures and properties of 134 kilo molecules](#). *Scientific data*, 1(1):1–7.
- B. Ramsundar, P. Eastman, P. Walters, and V. Pande. 2019. [Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More](#). O'Reilly.

- Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Dogan. 2018. [Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases](#). *Briefings in Bioinformatics*, 20:1878 – 1912.
- Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. [Get your atoms in order - an open-source implementation of a novel and robust molecular canonicalization algorithm](#). *Journal of chemical information and modeling*, 55 10:2111–20.
- Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Constantine Bekas, and Alpha Albert Lee. 2018. [Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction](#). *ACS Central Science*, 5:1572 – 1583.
- Hayal Bulbul Sonmez, Figen Kuloğlu, Koksal Karadag, and Fred Wudl. 2012. [Terephthalaldehyde- and isophthalaldehyde-based polyspiroacetals](#). *Polymer Journal*, 44:217–223.
- Hannes Stärk, D. Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Gunemann, and Pietro Lio'. 2021. [3d infomax improves gnn for molecular property prediction](#). In *International Conference on Machine Learning*.
- Bing Su, Dazhao Du, Zhao-Qing Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Haoran Sun, Zhiwu Lu, and Ji rong Wen. 2022. [A molecular multimodal foundation model associating molecule graphs with natural language](#). *ArXiv*, abs/2209.05481.
- Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. 2023. [Graphgpt: Graph instruction tuning for large language models](#). *ArXiv*, abs/2310.13023.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. [Galactica: A large language model for science](#). *ArXiv*, abs/2211.09085.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [Llama: Open and efficient foundation language models](#). *ArXiv*, abs/2302.13971.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *ArXiv*, abs/2307.09288.
- Ahmad Walid, Simon Elana, Chithrananda Seyone, Grand Gabriel, and Ramsundar Bharath. 2022. [Chemberta-2: Towards chemical foundation models](#). *arXiv preprint arXiv:2209.01712*.
- Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. [Smiles-bert: Large scale unsupervised pre-training for molecular property prediction](#). *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*.
- Wen Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Y. Qiao, and Jifeng Dai. 2023.

- Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *ArXiv*, abs/2305.11175.
- Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. 2024. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *arXiv preprint arXiv:2403.09513*.
- Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2021. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4:279 – 287.
- Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.
- Jinmao Wei, Xiao-Jie Yuan, Qinghua Hu, and Shuqin Wang. 2010. A novel measure for evaluating classifiers. *Expert Syst. Appl.*, 37:3799–3809.
- David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36.
- Aidan Wong, He Cao, Zijing Liu, and Yu Li. 2024. Smiles-prompting: A novel approach to llm jailbreak attacks in chemical synthesis. *arXiv preprint arXiv:2410.15641*.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Towards building open-source language models for medicine.
- Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. 2017. Moleculenet: A benchmark for molecular machine learning. *arXiv: Learning*.
- Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *ArXiv*, abs/2304.01196.
- Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *ArXiv*, abs/1810.00826.
- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. Gpt4tools: Teaching large language model to use tools via self-instruction. *ArXiv*, abs/2305.18752.
- Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phi Thi Mi, Haiquan Wang, Caiming Xiong, and Silvio Savarese. 2022. Retroformer: Pushing the limits of interpretable end-to-end retrosynthesis transformer. In *ICML*.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *ArXiv*, abs/2306.13549.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *ArXiv*, abs/2010.13902.
- Rong Yu, Bian Yatao, Xu Tingyang, Xie Weiyang, Wei Ying, Huang Wenbing, and Huang Junzhou. 2020. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571.
- Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, P. Zhang, Yuxiao Dong, and Jie Tang. 2022a. Glm-130b: An open bilingual pre-trained model. *ArXiv*, abs/2210.02414.
- Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022b. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature Communications*, 13.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning

for medical visual question answering. *ArXiv*, abs/2305.10415.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. [Uni-mol: A universal 3d molecular representation learning framework](#). In *International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. [Minigt-4: Enhancing vision-language understanding with advanced large language models](#). *ArXiv*, abs/2304.10592.



## A Tasks Definition and Dataset Details

**Property Prediction.** Molecular Property Prediction involves the forecasting or estimation of the biophysical and chemical properties of a molecule. In this work, our emphasis lies on three binary classification tasks sourced from the MoleculeNet benchmark (BBBP, BACE, and HIV) (Wu et al., 2017), and three regression tasks concentrating on the quantum properties of molecules from the QM9 (Ramakrishnan et al., 2014a) dataset.

**Molecule Description Generation.** Generating molecular descriptions involves compiling a detailed overview of a molecule’s structure, properties, activities, and functions. This process aids chemists and biologists by swiftly providing crucial molecular insights for their research. Our data collection involves the extraction of molecular text annotations from PubChem (Kim et al., 2022). Leveraging PubChem’s **Power User Gateway** (Kim et al., 2019), we retrieve abstracts of compound records in XML format. Subsequently, we extracted valid molecular description texts identified by unique PubChem Chemical Identifiers (CIDs), filtering out SMILES strings with syntactic errors or deviations from established chemical principles. Furthermore, we utilize the ChEBI-20 dataset (Edwards et al., 2021) for downstream tasks in molecule description generation, comprising 33,010 molecule description pairs divided into 80% for training, 10% for validation and 10% for testing. To prevent data leakage, compounds in the PubChem text annotations that coincide with the ChEBI-20 test split are excluded.

**Forward Reaction Prediction.** Predicting the forward reaction involves anticipating the probable product(s) of a chemical reaction based on given reactants and reagents. For this task, we utilize the forward-reaction-prediction dataset from (Fang et al., 2023), comprising 138,768 samples sourced from the USPTO dataset (Wei et al., 2010). Each entry includes reactants and reagents separated by ‘.’ within the instruction, with the output product.

**Reagent Prediction.** Reagent prediction identifies the substances necessary for a chemical reaction, helping to discover new types of reaction and optimal conditions. We use the reagent Prediction data from (Fang et al., 2023), sourced from the USPTO\_500MT dataset (Lu and Zhang, 2022b). Each entry features a chemical reaction indicated as

“reactants >> product,” with the output indicating the reagents involved in the reaction.

**Retrosynthesis Prediction.** Retrosynthetic analysis in organic chemistry reverses engineering by tracing potential synthesis routes from the target compound backward. This strategy is vital for the efficient synthesis of complex molecules and to foster innovation in pharmaceuticals and materials. For this task, we also used the dataset from (Fang et al., 2023), which is sourced from USPTO\_500MT. The data organize inputs as products and outputs as reactants separated by ‘.’ for each compound.

**Discussion on License.** As depicted in Table 6, we elaborate on the origins and legal permissions associated with each data component utilized in the development of the InstructMol. This encompasses both biomolecular data and textual descriptions. Thorough scrutiny was conducted on all data origins to confirm compatibility with our research objectives and subsequent utilization. Proper and accurate citation of these data sources is consistently maintained throughout the paper.

## B Implementation Details

**Model Settings.** A graph neural network with five graph isomorphism network (GIN) (Xu et al., 2018) layers is used as the molecule graph encoder  $f_g$ . The hidden dimension is set to be 300. The GIN model is initialized using the MoleculeSTM (Liu et al., 2022) graph encoder, which is pre-trained through molecular graph-text contrastive learning. We employ Vicuna-v-1.3-7B (Chiang et al., 2023) as the base LLM, which has been trained through instruction-tuning. The total number of parameters of InstructMol is around 6.9B.

**Training Details.** In the first stage, we employ the training split comprising around 264K molecule-caption pairs from PubMed. Using a batch size of 128, we conduct training for 5 epochs. We use the AdamW optimizer, with  $\beta=(0.9, 0.999)$  and a learning rate of  $2e-3$ , without weight decay. Warm-up is executed over 3% of the total training steps, followed by a cosine schedule for learning rate decay. For the second stage, we conduct training for three specific scenarios. For fair comparisons with traditional methods, training spans 20 to 50 epochs for the molecule description generation task using the ChEBI-20 training split. Property prediction and reaction tasks undergo 10 epochs using correspond-

TASKS	# SAMPLES	DATA SOURCE
Alignment Pretrain	264K	PubMed (Kim et al., 2022)
Property Prediction(Regression)	362K	QM9 (Fang et al., 2023; Wu et al., 2017)
Property Prediction(Classification)	35,742	BACE, BBBP, HIV (Wu et al., 2017)
Molecule Description Generation	26,507	ChEBI-20 (Edwards et al., 2021)
Forward Prediction	125K	USPTO (Fang et al., 2023; Wei et al., 2010)
Retrosynthesis	130K	USPTO_500MT (Fang et al., 2023; Lu and Zhang, 2022b)
Reagent Prediction	125K	USPTO_500K (Fang et al., 2023; Lu and Zhang, 2022b)

Table 5: Details of InstrutMol two-stage training data.

DATA SOURCES	LICENSE URL	LICENSE NOTE
PubChem	<a href="https://www.nlm.nih.gov/web_policies.html">https://www.nlm.nih.gov/web_policies.html</a>	Works produced by the U.S. government are not subject to copyright protection in the United States. Any such works found on National Library of Medicine (NLM) Web sites may be freely used or reproduced without permission in the U.S.
ChEBI	<a href="https://creativecommons.org/licenses/by/4.0/">https://creativecommons.org/licenses/by/4.0/</a>	You are free to: Share — copy and redistribute the material in any medium or format. Adapt — remix, transform, and build upon the material for any purpose, even commercially.
USPTO	<a href="https://www.uspto.gov/learning-and-resources/open-data-and-mobility">https://www.uspto.gov/learning-and-resources/open-data-and-mobility</a>	It can be freely used, reused, and redistributed by anyone.
MoleculeNet	<a href="https://opensource.org/license/mit/">https://opensource.org/license/mit/</a>	Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so.

Table 6: Data resources and licenses utilized in data collection..

ing instruction datasets. In InstructMol training, we maintain a consistent batch size of 128 and set the learning rate to  $8e-5$ . Linear layers within the LLM utilize a LoRA rank of 64 and a scaling value  $\alpha$  of 16. All experiments are run with  $4 \times$  RTX A6000 (48GB) GPUs.

Configuration	Value
Graph encoder $f_g$ init.	GIN <sub>MoleculeSTM</sub>
# params $f_g$	1.8M
LLM init.	Vicuna-v-1.3-7B
# params LLM	6.9B
Stage1 batch-size	128
Stage2 batch-size	128
Optimizer	AdamW
Warm-up ratios	0.03
Stage1 peak lr	$2e-3$
Stage2 peak lr	$8e-5$
Learning rate schedule	cosine decay
Weight decay	0.
Stage1 train epochs	5
Stage2 train epochs	20-50
Numerical precision	bfloat16
Activation checkpointing	True

Table 7: Training hyperparameters of InstructMol.

## C Evaluate Metrics

**Molecule Description Generation Metric.** Following (Edwards et al., 2022), NLP metrics such as BLEU (Papineni et al., 2001), ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) are used

to assess the proximity of generated descriptions to the truth of the ground. Specifically, these metrics are tested on the ChEBI-20 test dataset. In our experiments, we observed that after 50 epochs of finetuning on the training split, the metrics tend to converge, differing from previous approaches that often involved fine-tuning for over 100 epochs (Edwards et al., 2022; Su et al., 2022; Luo et al., 2023b).

**Molecule Generation Metric.** In chemical reaction tasks, we view it as akin to a text-based molecule generation task. Initially, we employ RDKit to validate the chemical validity of the generated results, ensuring their "validity". Subsequently, we gauge the sequential proximity between the generated sequence and the ground truth using NLP metrics such as BLEU, Exact Match scores, and Levenshtein distance. Additionally, we present performance based on molecule-specific metrics that assess molecular similarity, encompassing RDKit, MACCS (Durant et al., 2002), and Morgan (Schneider et al., 2015) fingerprints similarity.

TASK	INSTRUCTION
Alignment Pretrain	Instruction: <i>Provide a brief overview of this molecule.</i>    [Optional: The compound SELFIES sequence is: SELFIES] Output: <i>The molecule is a non-proteinogenic alpha-amino acid that is ...</i>
Property Prediction (Regression)	Instruction: <i>Could you give me the LUMO energy value of this molecule?</i>    [Optional: The compound SELFIES sequence is: SELFIES] Output: <i>0.0576</i>
Property Prediction (Classification)	Instruction: <i>Evaluate whether the given molecule is able to enter the blood-brain barrier.</i>    [Optional: The compound SELFIES sequence is: SELFIES] Output: <i>Yes</i>
Molecule Description Generation	Instruction: <i>Could you give me a brief overview of this molecule?</i>    [Optional: The compound SELFIES sequence is: SELFIES] Output: <i>The molecule is a fatty acid ester obtained by ...</i>
Forward Prediction	Instruction: <i>Based on the given reactants and reagents, suggest a possible product.</i>    <REACTANT A>.<REACTANT B>...<REAGENT A>.<REAGENT B>... Output: SELFIES of product
Retrosynthesis	Instruction: <i>Please suggest potential reactants used in the synthesis of the provided product.</i>    SELFIES of product Output: <REACTANT A>.<REACTANT B>...<REAGENT A>.<REAGENT B>...
Reagent Prediction	Instruction: <i>Can you provide potential reagents for the following chemical reaction?</i>    <REACTANT A>.<REACTANT B>...<REAGENT A>.<REAGENT B>... » <PRODUCTs> Output: SELFIES of reagent

Table 8: Examples of instruction samples for each task. || means concatenate along the token dimension.

```
messages = [ {"role": "system", "content": f"""You're acting as a molecule property prediction assistant. You'll be given SMILES of molecules and you need to make binary classification with a return result only in "True" or "False".
```

**The background of the dataset and task is shown below:**

The Blood-brain barrier penetration (BBBP) dataset comes from a recent study on the modeling and prediction of barrier permeability. As a membrane separating circulating blood and brain extracellular fluid, the blood-brain barrier blocks most drugs, hormones, and neurotransmitters. Thus penetration of the barrier forms a long-standing issue in the development of drugs targeting the central nervous system.

We provide several examples for this binary classification task:

###

Instruction: Predict whether the given compound has barrier permeability. Return True or False.

SMILES: CCC(=O)C(CC(C)N(C)C)(c1ccccc1)c2ccccc2

Output: True

###

###

Instruction: Predict whether the provided compound exhibits barrier permeability. Return True or False.

SMILES: c1cc2c(cc(CC3=CNC(=NC3=O)NCCSCc3oc(cc3)CN(C)C)cc2)cc1

Output: False

###

...

Given the following instructions and SMILES, return your prediction result:

Instruction: Predict whether the provided compound exhibits barrier permeability. Return True or False.

SMILES: TARGET SMILES

"""}

]

Table 9: An illustration of the few-shot in-context-learning prompt construction process for Llama (Touvron et al., 2023a,b) / Vicuna (Chiang et al., 2023) models in property prediction tasks.

## D More Results

### D.1 Ablation study results

METHODS	BLEU-2 $\uparrow$	BLEU-4 $\uparrow$	ROUGE-1 $\uparrow$	ROUGE-2 $\uparrow$	ROUGE-L $\uparrow$	METEOR $\uparrow$
<b>InstructMol-G</b>	0.4620	0.3560	0.5439	0.3644	0.4765	0.4832
+MLP XL connector	<b>0.4665(+0.97%)</b>	<b>0.3613(+1.49%)</b>	<b>0.5497(+1.07%)</b>	<b>0.3699(+1.51%)</b>	<b>0.4805(+0.84%)</b>	<b>0.4917(+1.76%)</b>
+Scale up LLM	0.4615(-0.11%)	0.3566(+0.17%)	0.5449(+0.18%)	0.3660(+0.44%)	0.4776(+0.23%)	0.4868(+0.75%)
Replace $f_g$ with GraphMVP	0.4452(-3.64%)	0.3377(-5.14%)	0.5318(-0.11%)	0.3484(-2.22%)	0.4638(-2.67%)	0.4691(-2.92%)
Skip Stage-1	0.4631(+0.23%)	0.3569(+0.25%)	0.5419(-0.37%)	0.3610(-0.93%)	0.4720(-0.94%)	0.4391(-9.12%)
Freeze LLM in the second stage	$\sim 0$	$\sim 0$	$\sim 0$	$\sim 0$	$\sim 0$	$\sim 0$

Table 10: Ablation of the model architecture and training scheme design. We chose to conduct experiments on the Molecule Description Generation task.  $f_g$  represents the molecule graph encoder.

### D.2 More Results of Molecule Description Generation


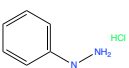
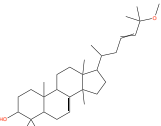
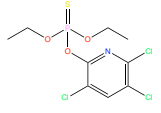
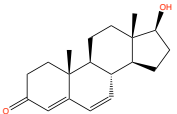
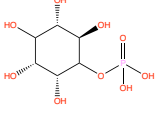
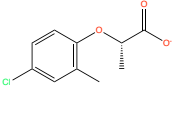
	Mol-Instruction	InstructMol	Ground Truth
 CID: 2735122	The molecule is the potassium salt of acetic acid. It has a role as a buffer. It is a potassium salt and an acetate salt. It contains an acetate.	The molecule is a potassium salt that is the monokis salt of carbonic acid. It has a role as a buffer. It is a carbonate salt and a potassium salt. It contains a carbonate.	The molecule is the potassium salt of formic acid. It has a role as a buffer. It is a potassium salt and a one-carbon compound. It derives from a formic acid.
 CID: 60962	The molecule is an organic molecular entity.	The molecule is a hydrochloride salt resulting from the reaction of equimolar amounts of 1-nitrobenzene and hydrogen chloride. It has a role as a mutagen. It contains a 1-nitrobenzene(2+).	The molecule is a hydrochloride resulting from the reaction of equimolar amounts of phenylhydrazine and hydrogen chloride. It contains a phenylhydrazine.
 CID: 50994221	The molecule is a triterpenoid.	The molecule is a tirucallane triterpenoid that is (13 $\alpha$ ,14 $\beta$ ,17 $\alpha$ ,20 $\beta$ ,23E)-lanosta-7,23-diene substituted by an oxo group at position 3 and a methoxy group at position 25. It has been isolated from the stem and stem barks of <i>Cornus walteri</i> . It has a role as a plant metabolite. It is an ether, a tirucallane triterpenoid and a secondary alcohol.	The molecule is a tirucallane triterpenoid that is (13 $\alpha$ ,14 $\beta$ ,17 $\alpha$ ,20 $\beta$ ,23E)-lanosta-7,23-diene substituted by a beta-hydroxy group at position 3 and a methoxy group at position 25. It has been isolated from the stem and stem barks of <i>Cornus walteri</i> . It has a role as a plant metabolite. It is a tirucallane triterpenoid, an ether and a secondary alcohol.
 CID: 2730	The molecule is an organic thiophosphate.	The molecule is an organic thiophosphate that is O,O-dimethyl hydrogen phosphorothioate in which the hydrogen of the hydroxy group has been replaced by a 3,5,6-trichloropyridin-2-yl group. It has a role as an agrochemical, an EC 3.1.1.7 (acetylcholinesterase) inhibitor and an environmental contaminant. It is an organic thiophosphate, a chloropyridine and an organochlorine pesticide.	The molecule is an organic thiophosphate that is O,O-diethyl hydrogen phosphorothioate in which the hydrogen of the hydroxy group has been replaced by a 3,5,6-trichloropyridin-2-yl group. It has a role as an EC 3.1.1.7 (acetylcholinesterase) inhibitor, an agrochemical, an EC 3.1.1.8 (cholinesterase) inhibitor, an environmental contaminant, a xenobiotic, an acaricide and an insecticide. It is an organic thiophosphate and a chloropyridine.
 CID: 17209	The molecule is a natural product found in <i>Homo sapiens</i> with data available.	The molecule is a 17 $\beta$ -hydroxy steroid that is testosterone in which the 17 $\beta$ hydrogen is replaced by a methyl group. It has a role as an androgen. It is a 3-oxo-Delta(4) steroid and a 17 $\beta$ -hydroxy steroid. It derives from a testosterone.	The molecule is a 17 $\beta$ -hydroxy steroid that is testosterone that contains an additional double bond between positions 6 and 7. It is a 3-oxo-Delta(4) steroid, a 17 $\beta$ -hydroxy steroid and an enone. It derives from a testosterone.
 CID: 107737	The molecule is a metabolite found in or produced by <i>Escherichia coli</i> (strain K12, MG1655).	The molecule is a myo-inositol monophosphate. It has a role as an algal metabolite and a mouse metabolite. It derives from a myo-inositol. It is a conjugate acid of a 1D-myo-inositol 4-phosphate(2-).	The molecule is an inositol having myo-configuration substituted at position 1 by a phosphate group. It has a role as a human metabolite, an <i>Escherichia coli</i> metabolite and a mouse metabolite. It derives from a myo-inositol. It is a conjugate acid of a 1D-myo-inositol 1-phosphate(2-).
 CID: 107737	The molecule is a monocarboxylic acid anion resulting from the removal of a proton from the carboxy group of (R)-imazamox. It is a conjugate base of a (R)-imazamox(1-)	The molecule is a monocarboxylic acid anion resulting from the removal of a proton from the carboxy group of (S)-methyl 2-(4-chloro-2-methylphenoxy)acetate. It is a conjugate base of a (S)-methyl 2-(4-chloro-2-methylphenoxy)acetate. It is an enantiomer of a (R)-methyl 2-(4-chloro-2-methylphenoxy)acetate(1-).	The molecule is a monocarboxylic acid anion that is the conjugate base of (S)-2-(4-chloro-2-methylphenoxy)propanoic acid, obtained by deprotonation of the carboxy group. It is a conjugate base of a (S)-mecoprop. It is an enantiomer of a (R)-2-(4-chloro-2-methylphenoxy)propanoate.

Figure 4: More examples of molecule description generation task on ChEBI-20 (Edwards et al., 2021) test set. We include Mol-Instruction (Fang et al., 2023) as the baseline. CID (CID): PubChem Compound Identification, a non-zero integer PubChem accession identifier for a unique chemical structure.

### D.3 More Results of Forward Reaction Prediction

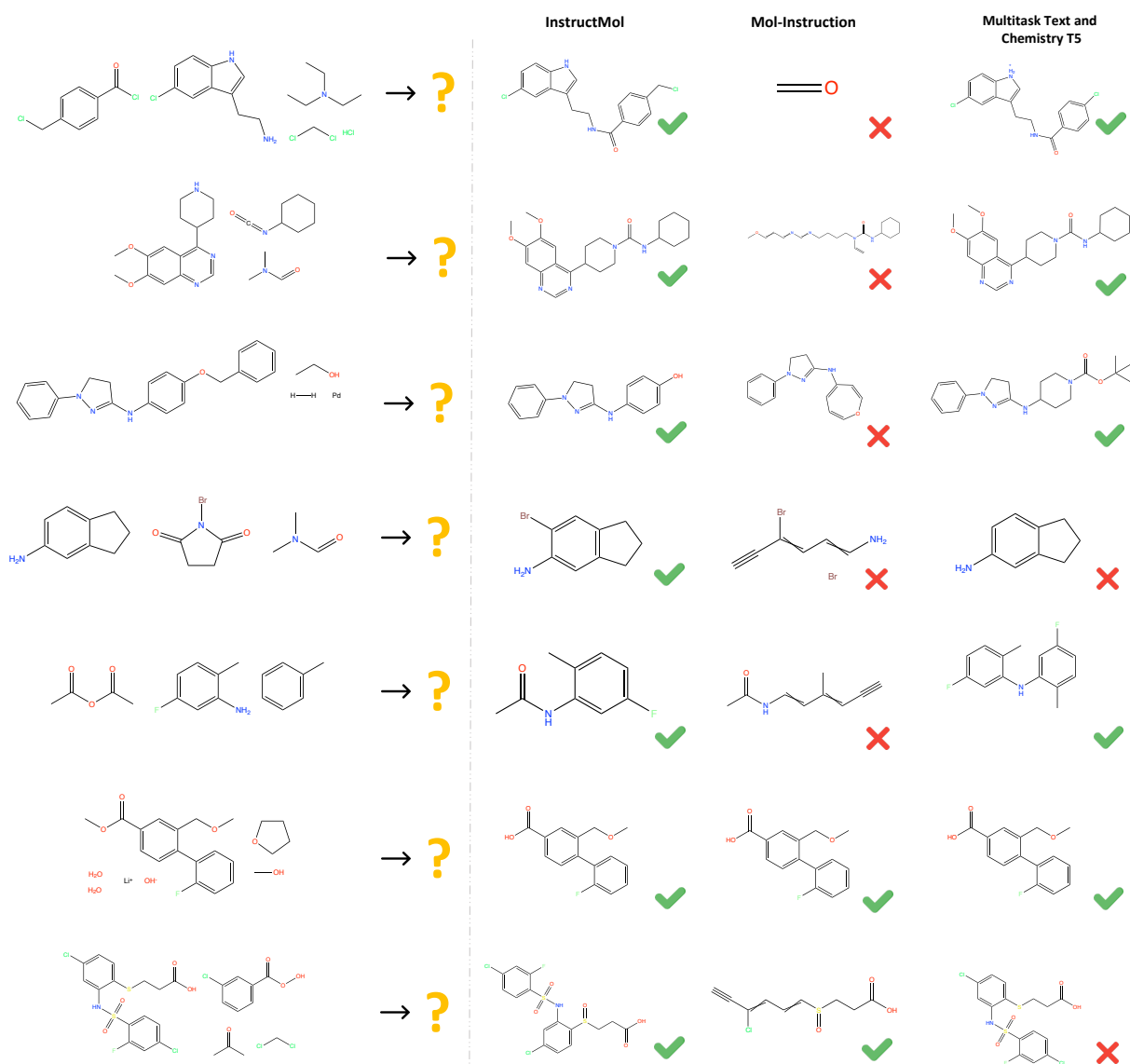


Figure 5: More examples of forward reaction prediction task. We include Mol-Instruction (Fang et al., 2023) and Multitask-Text-and-Chemistry-T5 (Christofidellis et al., 2023) as baselines.

## D.4 More Results of Reagent Prediction

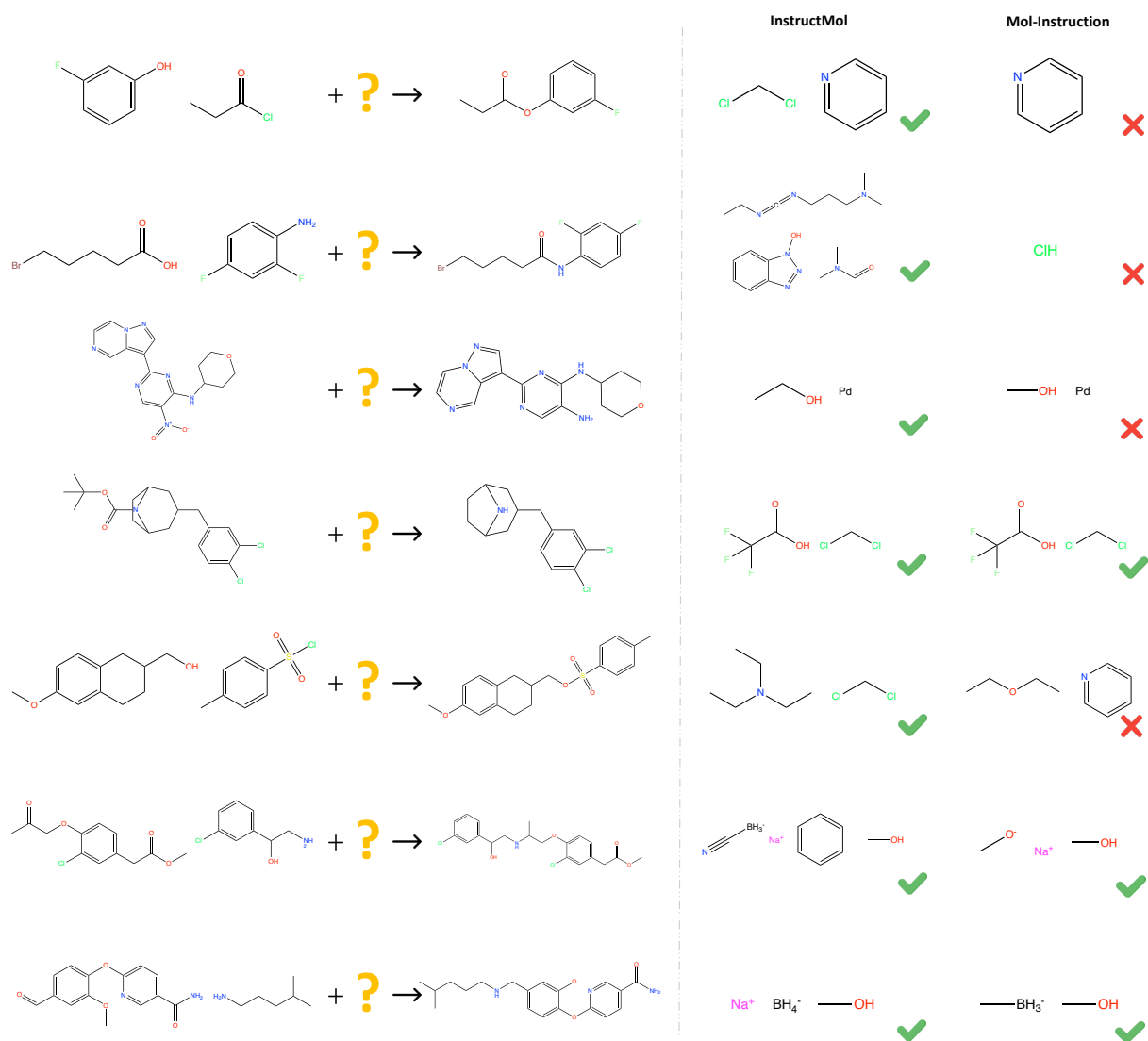


Figure 6: More examples of the reagent prediction task. We include Mol-Instruction (Fang et al., 2023) as the baseline.

## D.5 More Results of Retrosynthesis Prediction

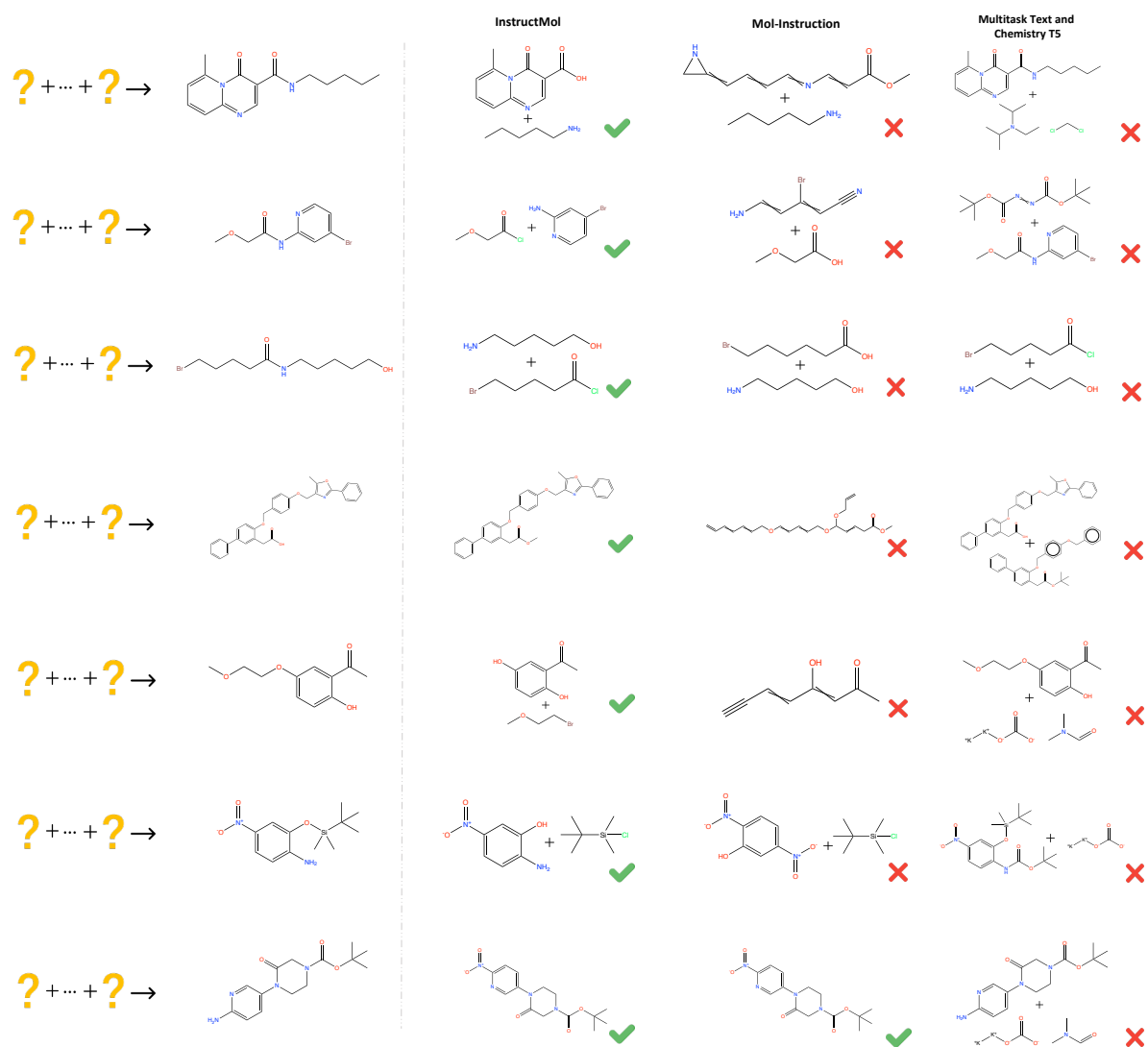


Figure 7: More examples of the retrosynthesis prediction task. We include Mol-Instruction (Fang et al., 2023) and Multitask-Text-and-Chemistry-T5 (Christofidellis et al., 2023) as baselines.

## D.6 Error Analysis

We showcase cases with misalignment to the ground truth, along with RDKit fingerprint similarity results in Fig. 8. The complexity of chemical reaction compounds makes the task more challenging. In addressing this limitation, our future approach involves concatenating graph tokens from multiple molecules involved in the same reaction with text tokens to simplify the complexity of the input sequence. Moreover, we are considering employing separate tokenization and embedding for distinct modalities to ensure the semantic accuracy of the tokenized results.

## D.7 More Results of Molecule Property Prediction

Based on the molecular property binary classification discussed in the main text, we have extended our analysis by comparing Instruct-G with other domain-specific models (ChemBERTa (Chithrananda et al., 2020), GROVER (Yu et al., 2020), DMP (Jinhua et al., 2023)) and MolCA variants on multi-class prediction tasks from MoleculeNet, including Clintox, Tox21, Toxcast, and SIDER, as detailed in the Table 11.

METHOD	Clintox	Tox21	Toxcast	SIDER
# MOLECULES	1491	8014	8615	1427
# TASKS	2	12	617	27
<i>Specialist Models</i>				
ChemBERTa (Chithrananda et al., 2020)	73.3	-	-	-
ChemBERTa2 (Walid et al., 2022)	23.9	-	-	-
GROVER-large (Yu et al., 2020)	94.4	83.1	73.7	65.8
DMP(TF+GNN) (Jinhua et al., 2023)	95.6	79.1	-	69.8
MolCA(1D+2D) (Liu et al., 2023f)	89.5	77.2	64.5	-
<b>Instruct-G</b>	93.4	77.0	61.9	62.2

Table 11: ROC-AUC results of molecular property tasks (multi-classes classification) on MoleculeNet (Wu et al., 2017) benchmarks.

Based on the results of Table 2 and Table 11, we found that integrating the 1D and 2D molecular modalities significantly enhances the model’s understanding capabilities. It is important to note that ChemBERTa, GROVER, and DMP were all pre-trained on large molecule-only datasets: ChemBERTa on 77M unique SMILES, GROVER on 11M molecules, and DMP on 110M molecules. In contrast, InstructMol utilized only about 300K molecule-text description pairs for the initial alignment stage, with parameter size updates confined to the projector layer ( $< 1$  million), and without extensive retraining. This limited the molecule space it covered. To further improve performance on MoleculeNet, additional pretraining stages and

the collection of large unlabeled datasets to cover a broader range of molecules could be considered.

## D.8 From LoRA to Full-Finetuning

InstructMol is instruction-tuned using LoRA, with a trainable parameter size of less than 100M, which is significantly lower than that of domain expert models like the MolT5 (Edwards et al., 2022) series. These domain expert models are pretrained on over 100 million SMILES and are limited to only a few tasks, such as molecule captioning and de novo design. The main focus of our work is to demonstrate that the aligning SFT training approach can efficiently and rapidly adapt general language models into domain-specific multimodal models capable of addressing multiple downstream tasks. Increasing the trainable parameters and adding additional pre-training datasets will further boost InstructMol’s performance, as shown in Table 12.

To assess whether InstructMol can retain the original capabilities of LLMs, we conducted additional dialogues using InstructMol. Our findings indicate that the model continues to exhibit communication skills, common sense inference, and logical reasoning at a qualitative level. Additionally, we provided quantitative results on several MMLU tasks (Hendrycks et al., 2020) (zero-shot) in Table 13, demonstrating that despite the inevitable forgetting problem introduced by fine-tuning, InstructMol retains most of the original LLM’s capabilities.



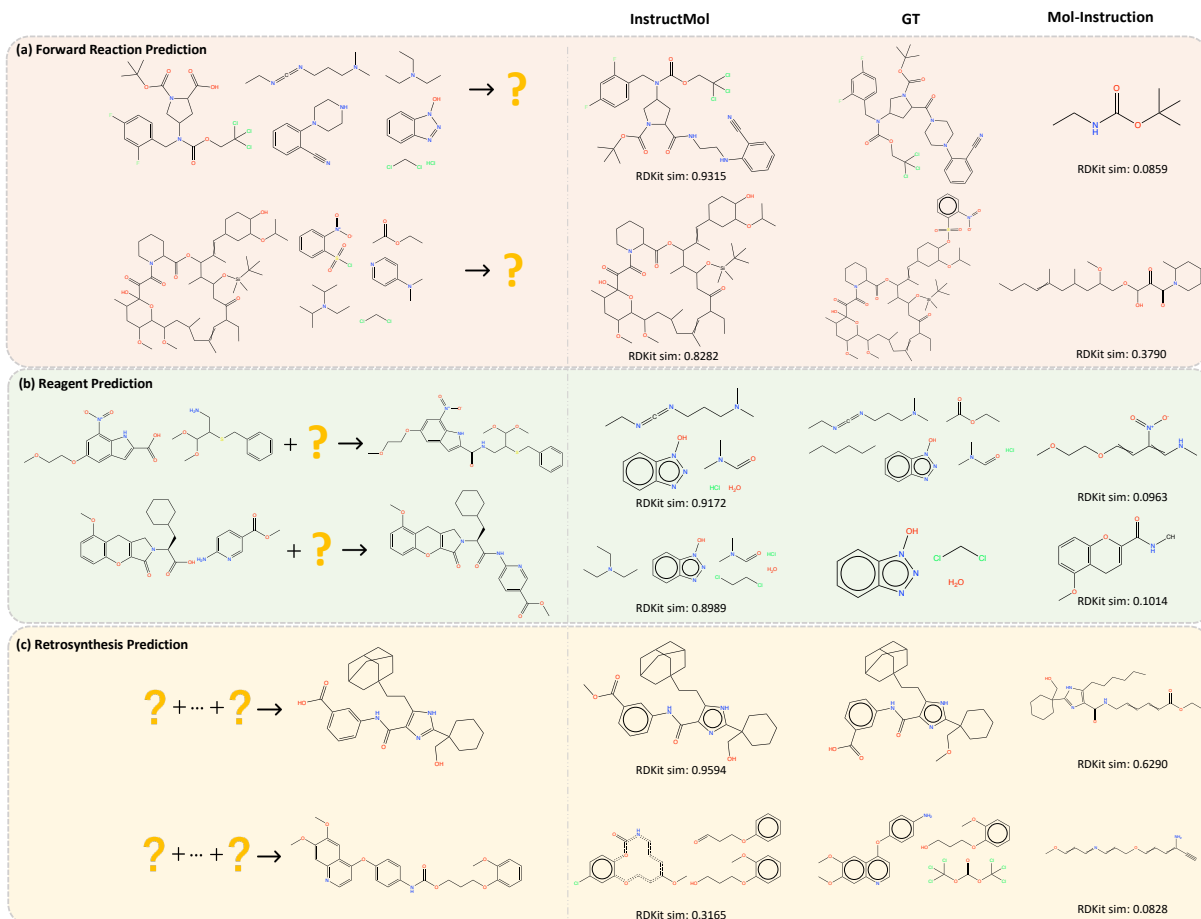


Figure 8: We present several cases with a certain degree of misalignment compared to the ground truth, accompanied by RDKit fingerprint similarity results relative to the ground truth. Due to the heightened complexity of compounds involved in chemical reactions, the difficulty of the task increases, leading to the poor performance of Mol-Instructions (Fang et al., 2023).

METHODS	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
MolT5-large (Edwards et al., 2022)	0.594	0.508	0.654	0.510	0.594	0.614
Text+Chem T5 (base) (Christofidellis et al., 2023)	0.625	0.542	0.682	0.543	0.622	0.648
MolCA (Liu et al., 2023f)	0.620	0.531	0.681	0.537	0.618	0.651
<b>Instruct-G (Full-tune)</b>	0.653	0.566	0.608	0.445	0.541	0.562

Table 12: Comparison with state-of-the-art models on the Molecule Caption task when performing full fine-tuning.

Model	High School			College		
	Biology	Physics	Chemistry	Biology	Physics	Chemistry
Vicuna-7B-v1.3	0.529	0.291	0.345	0.465	0.186	0.270
InstructMol-GS	0.481	0.258	0.246	0.438	0.196	0.230

Table 13: Performance comparison of LoRA-tuned models with original models across MMLU high school and college science subjects.

## E Comparison with Current Agents Framework

LLMs face a major limitation in performing basic mathematical and chemical operations, which makes handling hallucinations challenging. However, their self-supervised pre-training on diverse knowledge equips them with a strong understanding and reasoning abilities that can be directly applied to new domains. Presenting LLMs as automated assistants offers a programming-free interface for non-experts to leverage their existing capabilities. Agent/assistant paradigms enable the optimal utilization of LLMs' knowledge without the need for specialized model development. For instance, ChemCrow (Bran et al., 2023) is an agent system based on GPT-4 that integrates various chemical tools for solving diverse tasks. We conducted a comparison of three downstream tasks between InstructMol and ChemCrow, and the results are presented in Table 14.

During testing, we observed that ChemCrow's performance is heavily reliant on prompt construction, resulting in unstable output results. For instance, in retrosynthesis planning experiments, we found that agents often misidentify the user's query product as controlled chemistry and refuse to provide an answer. Similarly, in the property prediction task, GPT-4 itself lacks specific knowledge about compounds and thus heavily relies on internet searches. The quality of the prompt constructed by the user significantly influences the quality of the response.

Task	Ground Truth	ChemCrow	InstructMol
<i>Property Prediction</i>			
Determine whether (CID:219214) can suppress HIV.	"Active"	WebSearch→ No information	✓
<i>Forward Reaction Prediction</i>			
<chem>CCC(=O)Cl + OC1=CC=CC(F)=C1 + ClCCl + C2=CC=NC=C2</chem> →?	<chem>CCC(=O)OC1=CC=CC(F)=C1</chem>	✓	✓
<i>Retrosynthesis Prediction</i>			
? → <chem>C(CCNC(=O)CCCCBr)CCO</chem>	<chem>NCCCCCO.O=C(O)CCCCBr</chem>	"Similar to controlled chemistry, reject to answer"	✓

Table 14: The performance of InstructMol and ChemCrow was evaluated through a comparison of three downstream tasks: Property Prediction, Forward Reaction Prediction, and Retrosynthesis. The ✓ denotes that the predictions match with the ground truths.

Therefore, we believe that domain-specific LLMs should be augmented with dedicated external tools. This augmentation would enable LLMs to function as planners, comprehend and decompose tasks, invoke downstream interfaces, and effectively process feedback. In our future work, we intend to create a new dataset for instruction-following tool usage and enhance InstructMol with a variety of external tools. By leveraging state-of-the-art models and maximizing LLM's reasoning and planning capabilities, we aim to further enhance its performance.