

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

Anonymous ACL submission

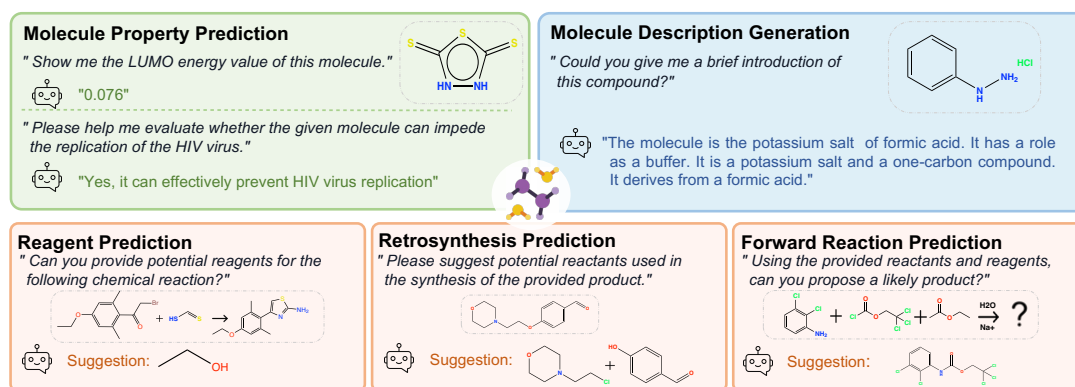


Figure 1: Empowering LLMs with molecular modalities to unlock the drug discovery domain and serve as assistants in molecular research.

## 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**, 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) and helps to understand the complex molecular space, reducing iterative testing and minimizing failure rates. Previous approaches involved employ-

ing 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. Through a unified model, it 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

060	activities;		
061	• Developing an effective training paradigm that		
062	guides LLMs in utilizing molecular representa-		
063	tions and adapting to various tasks.		
064	Several prior studies (Liang et al., 2023; Luo et al.,		
065	2023c; Fang et al., 2023) have fine-tuned generalist		
066	LLMs to develop foundational models within the		
067	molecular domain. Despite their enhancement to		
068	the original generalist LLM, these preceding works		
069	have unveiled several <b>issues</b> :		
070	1. Insufficient alignment between different modal-		
071	ities.		
072	2. The consideration of an optimal molecular struc-		
073	ture encoder remains unexplored.		
074	3. A rudimentary design of the training pipeline		
075	neglects the update of LLMs’ knowledge.		
076	These extant issues lead to a significant dispar-		
077	ity in the performance of the current AI assistants		
078	across various practical tasks when compared to		
079	traditional specialist models.		
080	To address these problems, we introduce <b>In-</b>		
081	<b>structMol</b> (Figure 2), a multi-modality instruction-		
082	tuning-based LLM. This model aligns molecular		
083	graphs and chemical sequential modalities with		
084	the natural language of humans. Using a cali-		
085	brated collection of molecule-related instruction		
086	datasets and a two-stage training scheme, <b>Instruct-</b>		
087	<b>Mol</b> effectively leverages the pre-trained LLM and		
088	molecule graph encoder for molecule-text align-		
089	ment. In the first alignment pretraining stage,		
090	we employ molecule-description pairs to train a		
091	lightweight and adaptable interface, which is de-		
092	signed to project the molecular node-level repre-		
093	sentation into the textual space that the LLM can		
094	understand. Subsequently, we finetune with multi-		
095	ple task-specific instructions. During this process,		
096	we freeze the molecule graph encoder and train		
097	low-rank adapters (LoRA (Hu et al., 2021)) on the		
098	LLM to adapt our model to various scenarios. This		
099	efficient approach enables the seamless integration		
100	of molecular and textual information, promoting		
101	the development of versatile and robust cognitive		
102	abilities in the molecular domain.		
103	To illustrate the capabilities of our model, we		
104	perform experiments that span three facets of drug		
105	discovery-related tasks, including compound prop-		
106	erty prediction, molecule description generation,		
107	and analysis of chemical reactions involving com-		
108	pounds. These tasks serve as robust benchmarks		
109	to assess the model’s ability to deliver useful and		
110	accurate knowledge feedback in practical drug dis-		
111	covery scenarios. The results in all experiments		
	consistently indicate that our model significantly		
	improves the performance of LLMs in tasks re-		
	lated 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 <b>InstructMol</b> , 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 an-		
	notated data in the drug discovery domain, our		
	approach strives to efficiently extract molecular		
	representations (targets on <b>Issue2</b> ). Employing a		
	two-stage instruction tuning paradigm enhances		
	the LLM’s understanding of molecular structural		
	and sequential knowledge (targets on <b>Issue1</b> and		
	<b>Issue3</b> ).		
	• InstructMol enables swift fine-tuning, generat-		
	ing lightweight checkpoints (used as plugins) for		
	cross-modality tasks. It provides the flexibility		
	to load or combine functionalities through plu-		
	gins, retaining the open dialogue and reasoning		
	capabilities of a general LLM.		
	• We evaluate our model through multiple prac-		
	tical assessments, demonstrating its substantial		
	improvement compared to state-of-the-art LLMs.		
	Our work lays the foundation for creating a ver-		
	satile and reliable molecular research assistant in		
	the drug discovery domain.		
	<b>2 Related Work</b>		
	<b>2.1 Multimodal Instruction Tuning</b>		
	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 instruc-		
	tion tuning, where pre-trained LLMs are fine-tuned		
	on instruction-formatted datasets (Wei et al., 2021),		
	allowing them to generalize to new tasks. Re-		
	cently, with the emergence of large foundation mod-		
	els in various domains, several efforts have been		
	made to transition from unimodal LLMs to multi-		
	modal 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 in-		

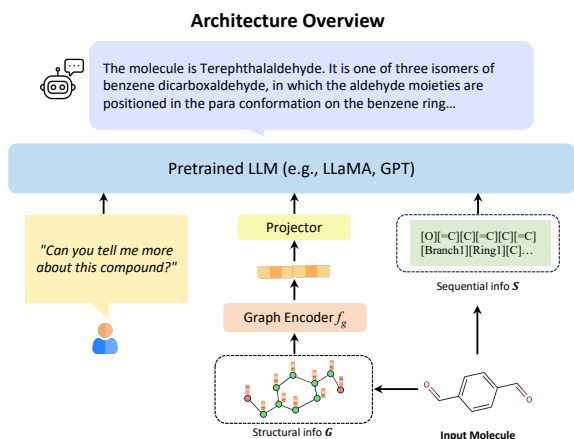


Figure 2: Overview of **InstructMol** model architecture design and two-stage training paradigm. The example molecule in the figure is *Terephthalaldehyde* (Sonmez et al., 2012) (CID 12173).

struction 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

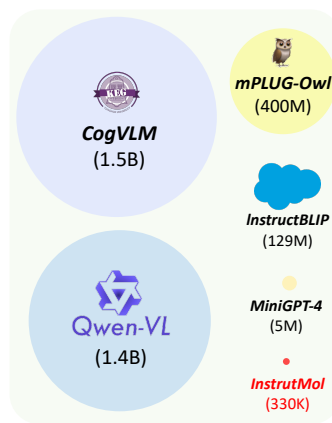
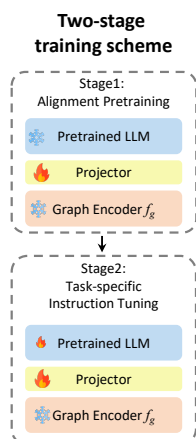


Figure 3: Comparison of biomolecule-domain molecule-text dataset scale with existing general domain vision-language datasets.

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, accurately 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). 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).

**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.

### 229 3 Method

#### 230 3.1 Multimodal Instruction Tuning

231 Instruction tuning refers to finetuning pretrained  
232 LLMs on instruction datasets, enabling generaliza-  
233 tion to specific tasks by adhering to new instruc-  
234 tions. Multimodal instruction tuning integrates  
235 modalities like images and graphs into an LLM,  
236 expanding the model’s capability to accommodate  
237 multiple modalities.

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

#### 253 3.2 Construction of Molecular Instruction

254 **Data Collection.** In the field of biomolecular re-  
255 search, there is a noticeable scarcity of molecular  
256 datasets with comprehensive text annotations when  
257 compared to the vision-language domain, as de-  
258 picted in Figure 3. While it is possible to construct  
259 instruction datasets in general domains by adapting  
260 benchmarks or using self-instruction, the applica-  
261 tion of these methods in the biomolecular domain  
262 presents challenges. This difficulty arises from two  
263 main factors: 1) biomolecular domain annotation  
264 demands expert knowledge and entails substantial  
265 complexity; 2) the knowledge within this domain  
266 spans a broad range of subjects, including struc-  
267 tural biology, computational chemistry, and chemi-  
268 cal synthesis processes.

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

275 **Molecule Input.** We utilize both the structure and  
276 sequence information of a molecule. We encode  
277 the structural information of a molecule as a graph,  
278 denoted by  $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}, \mathcal{X})$ , where  $\mathcal{V}$  is the set

279 of atoms (nodes) and  $|\mathcal{V}| = N$  is the total number  
280 of atoms. The set of edges  $\mathcal{E}$  includes all chemical  
281 bonds, and  $\mathcal{A} \in \mathbb{R}^{N \times N}$  is the adjacency matrix.  
282 Additionally,  $\mathcal{X} \in \mathbb{R}^{N \times F}$  encompasses attributes  
283 associated with each node, where  $F$  is the feature  
284 dimension. With a Graph Encoder  $f_g$ , we extract a  
285 graph representation  $\mathbf{Z}_G \in \mathbb{R}^{N \times d}$  at the node level,  
286 effectively describing the inherent structure of the  
287 molecule. Simultaneously, we consider encoding  
288 the sequential information of the molecule, denoted  
289 as  $S$ , as a supplementary source of structural infor-  
290 mation. To enhance the robustness of sequential  
291 molecular descriptors and mitigate syntactic and  
292 semantic invalidity present in SMILES (Weininger,  
293 1988), we employ SELFIES (Krenn et al., 2019) as  
294  $S$ , which is designed for mapping each token to a  
295 distinct structure or reference.

296 **Input Formulation.** We formulate a molecule-text  
297 pair ( $\mathbf{X}_M$  &  $\mathbf{X}_I$ ) to the corresponding instruction-  
298 following version like Human:  $\mathbf{X}_I \langle \text{mol} \rangle \mathbf{X}_M$   
299  $\langle \text{STOP} \rangle$  Assistant:  $\mathbf{X}_A \langle \text{STOP} \rangle$ . The  $\mathbf{X}_M$  repre-  
300 sents the molecule, including the molecule graph  
301  $\mathbf{X}_G$  and optionally the SELFIES  $\mathbf{X}_S$ .  $\mathbf{X}_I$  de-  
302 notes for the instruction and  $\mathbf{X}_A$  is the answer. For  
303 a given answer sequence of length  $L$ , our optimiza-  
304 tion objective is to maximize the probability of  
305 generating the target answers  $X_A$  by maximizing:

$$254 \quad p(\mathbf{X}_A | \mathbf{X}_M, \mathbf{X}_I) = \prod_{i=1}^L p_{\theta}(x_i | \mathbf{X}_G \parallel \mathbf{X}_S, \mathbf{X}_I, \mathbf{X}_{A, <i}). \quad (1)$$

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

#### 315 3.3 Architecture

316 **Molecular Encoder.** The molecular graph-  
317 structure encoder,  $f_g$ , needs to effectively extract  
318 node representations while preserving the molecu-  
319 lar graph’s connectivity information. It is crucial  
320 that  $f_g$  inherently establishes a pre-alignment in the  
321 representation space with the text space to facilitate  
322  $\mathbf{Z}_G$  in the following alignment stage. Taking inspi-  
323 ration from common practices in the Vision Large  
324 Language Models (VLLM) domain (Bai et al.,  
325 2023; Liu et al., 2023b; Ye et al., 2023), where  
326 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.** In order 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 the training phase, 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 the training process, we utilize the parameters of the alignment projector that were 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 adaptation allows the LLM to retain the inherent capacity for common-sense reasoning in dialogue.

## 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. 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 HUMO, LUMO, and the HUMO-LUMO gap (Ramakrishnan et al., 2014b). For the classification task, we incorporate three binary classification datasets pertaining to 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 (Ram-sundar 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) for QM9 property prediction regression tasks. †: few-shot in-context learning (ICL) results from Fang et al. (2023).  $\Delta\epsilon$ : HOMO-LUMO energy gap.

**Results.** Our models are compared against baselines on the test set for regression, measured by

METHOD # MOLECULES	BACE ↑ 1513	BBBP ↑ 2039	HIV ↑ 41127
<i>Specialist Models</i>			
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	<b>72.9</b>	78.8
Uni-Mol (Zhou et al., 2023)	<b>85.7</b>	<b>72.9</b>	<b>80.8</b>
<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
<b>Instruct-G</b>	<b>84.3</b> ( $\pm 0.6$ )	68.6 ( $\pm 0.3$ )	<b>74.0</b> ( $\pm 0.1$ )
<b>Instruct-GS</b>	82.1 ( $\pm 0.1$ )	<b>72.4</b> ( $\pm 0.3$ )	68.9 ( $\pm 0.3$ )

Table 2: ROC-AUC results of molecular property prediction tasks (classification) on MoleculeNet (Wu et al., 2017) benchmarks. \*: use LoRA tuning.

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 classification 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 depiction of a molecule, 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 (Ed-

MODEL	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
<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	<b>0.651</b>
Text+Chem T5-augm-base (Christofidellis et al., 2023)	<b>0.625</b>	<b>0.542</b>	<b>0.682</b>	<b>0.543</b>	<b>0.622</b>	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.

wards 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 multi-modalities (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 Prediction*: ascertain the suitable catalysts, solvents, or ancillary substances required for a specific chem-

ical 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 subsequently 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 by a significant margin. 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

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	0.0.673	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>Retroanalysis</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)	0.536	0.881	10.277	0.865	0.904	0.830	0.995
<b>InstructMol-G</b>	0.114	0.586	21.271	0.422	0.523	0.285	1.000
<b>InstructMol-GS</b>	<b>0.407</b>	<b>0.941</b>	<b>13.967</b>	<b>0.753</b>	<b>0.852</b>	<b>0.714</b>	1.000

Table 4: Results of chemical reaction tasks. These tasks encompass reagent prediction, forward reaction prediction, and retrosynthesis.  $\dagger$ : few-shot ICL results from (Fang et al., 2023). \*: use task-specific instruction data to finetune.

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) 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.



## 629 Limitations

630 In our investigation, several limitations have  
631 emerged. Firstly, the scale and quality of the  
632 dataset pose significant constraints; the scarcity  
633 of high-quality annotated domain data may hinder  
634 the model’s ability to generalize across the diverse  
635 and intricate molecular landscapes encountered in  
636 real-world applications. Secondly, the integration  
637 and evaluation of multiple modalities have also re-  
638 vealed areas needing improvement. Further refine-  
639 ment is necessary to ensure robust alignment and  
640 utilization of different molecule modalities within  
641 the model, enhancing its capacity to interpret and  
642 generate responses accurately across the molecular  
643 domain. Lastly, our base LLM originates from a  
644 general-domain model. However, the absence of  
645 specialized LLMs tailored specifically for chem-  
646 istry and molecular science, like models such as  
647 LLaMA, highlights the need for larger, more versa-  
648 tile domain-specific LLMs to enhance performance  
649 and expand applications. Addressing these chal-  
650 lenges is pivotal for enhancing the model’s reli-  
651 ability and extending its utility in advancing drug  
652 discovery methodologies.

## 653 Potential Risks

654 The application of AI in drug discovery entails  
655 several potential risks. A primary concern is the po-  
656 tential misuse of AI to develop hazardous or illicit  
657 substances, which presents significant safety and  
658 ethical challenges. Moreover, inaccuracies in AI-  
659 generated outputs could lead to hazardous chemical  
660 reactions if not thoroughly verified, posing risks  
661 of harm or damage to equipment. Dependence  
662 on AI-generated content heightens the risk of ac-  
663 cidents and unsafe practices. Therefore, stringent  
664 oversight and rigorous adherence to ethical guide-  
665 lines are essential to mitigate these risks and ensure  
666 the safe and responsible application of AI in drug  
667 discovery.

## 668 References

- 669 PubChem Structure Search. [https://pubchem.ncbi.nlm.nih.gov/search/  
670 help\\_search.html](https://pubchem.ncbi.nlm.nih.gov/search/help_search.html).  
671
- 672 Jean-Baptiste Alayrac, Jeff Donahue, Pauline  
673 Luc, Antoine Miech, Iain Barr, Yana Has-  
674 son, Karel Lenc, Arthur Mensch, Katie Mil-  
675 lican, Malcolm Reynolds, Roman Ring, Eliza  
676 Rutherford, Serkan Cabi, Tengda Han, Zhitao

Gong, Sina Samangooei, Marianne Monteiro, Ja- 677  
cob Menick, Sebastian Borgeaud, Andy Brock, 678  
Aida Nematzadeh, Sahand Sharifzadeh, Miko- 679  
laj Binkowski, Ricardo Barreira, Oriol Vinyals, 680  
Andrew Zisserman, and Karen Simonyan. 2022. 681  
*Flamingo: a visual language model for few-shot* 682  
*learning*. *ArXiv*, abs/2204.14198. 683

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin 684  
Johnson, Dmitry Lepikhin, Alexandre Tachard 685  
Passos, Siamak Shakeri, Emanuel Taropa, 686  
Paige Bailey, Z. Chen, Eric Chu, J. Clark, 687  
Laurent El Shafey, Yanping Huang, Kath- 688  
leen S. Meier-Hellstern, Gaurav Mishra, Er- 689  
ica Moreira, Mark Omernick, Kevin Robin- 690  
son, Sebastian Ruder, Yi Tay, Kefan Xiao, 691  
Yuanzhong Xu, Yujing Zhang, Gustavo Hernán- 692  
dez Abrego, Junwhan Ahn, Jacob Austin, 693  
Paul Barham, Jan A. Botha, James Brad- 694  
bury, Siddhartha Brahma, Kevin Michael 695  
Brooks, Michele Catasta, Yongzhou Cheng, 696  
Colin Cherry, Christopher A. Choquette-Choo, 697  
Aakanksha Chowdhery, C Crépy, Shachi Dave, 698  
Mostafa Dehghani, Sunipa Dev, Jacob Devlin, 699  
M. C. D’iaz, Nan Du, Ethan Dyer, Vladimir 700  
Feinberg, Fan Feng, Vlad Fienber, Markus Fre- 701  
itag, Xavier García, Sebastian Gehrmann, Lu- 702  
cas González, Guy Gur-Ari, Steven Hand, Hadi 703  
Hashemi, Le Hou, Joshua Howland, An Ren 704  
Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, 705  
Abe Ittycheriah, Matthew Jagielski, Wen Hao 706  
Jia, Kathleen Kenealy, Maxim Krikun, Sneha 707  
Kudugunta, Chang Lan, Katherine Lee, Ben- 708  
jamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang 709  
Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong- 710  
Zhong Liu, Frederick Liu, Marcello Maggioni, 711  
Aroma Mahendru, Joshua Maynez, Vedant 712  
Misra, Maysam Moussalem, Zachary Nado, 713  
John Nham, Eric Ni, Andrew Nystrom, Alicia 714  
Parrish, Marie Pellat, Martin Polacek, Oleksandr 715  
Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, 716  
Bryan Richter, Parker Riley, Alexandra Ros, 717  
Aurko Roy, Brennan Saeta, Rajkumar Samuel, 718  
Renee Marie Shelby, Ambrose Slone, Daniel 719  
Smilkov, David R. So, Daniela Sohn, Simon 720  
Tokumine, Dasha Valter, Vijay Vasudevan, Kiran 721  
Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui 722  
Wang, Tao Wang, John Wieting, Yuhuai Wu, 723  
Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng 724  
Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, 725  
Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, 726  
and Yonghui Wu. 2023. *Palm 2 technical report*. 727

728	<a href="#">ArXiv, abs/2305.10403.</a>	<a href="#">Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.</a>	775
729	Heba Askr, Enas Elgeldawi, Heba Aboul Ella,	Seyone Chithrananda, Gabriel Grand, and Bharath	777
730	Yaseen A.M.M. Elshaier, Mamdouh M. Gomaa,	Ramsundar. 2020. <a href="#">Chemberta: Large-scale self-</a>	778
731	and Aboul Ella Hassanien. 2022. <a href="#">Deep learning</a>	<a href="#">supervised pretraining for molecular property</a>	779
732	<a href="#">in drug discovery: an integrative review and fu-</a>	<a href="#">prediction.</a> <i>ArXiv, abs/2010.09885.</i>	780
733	<a href="#">ture challenges.</a> <i>Artificial Intelligence Review,</i>	Dimitrios Christofidellis, Giorgio Giannone, Jannis	781
734	<i>56:5975 – 6037.</i>	Born, Ole Winther, Teodoro Laino, and Mat-	782
735	Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang,	teo Manica. 2023. <a href="#">Unifying molecular and tex-</a>	783
736	Sinan Tan, Peng Wang, Junyang Lin, Chang	<a href="#">tual representations via multi-task language mod-</a>	784
737	Zhou, and Jingren Zhou. 2023. <a href="#">Qwen-vl: A</a>	<a href="#">elling.</a> In <i>International Conference on Machine</i>	785
738	<a href="#">frontier large vision-language model with versa-</a>	<i>Learning.</i>	786
739	<a href="#">tile abilities.</a> <i>ArXiv, abs/2308.12966.</i>	Connor W. Coley. 2020. <a href="#">Defining and exploring</a>	787
740	Satanjeev Banerjee and Alon Lavie. 2005. <a href="#">Meteor:</a>	<a href="#">chemical spaces.</a> <i>Trends in Chemistry.</i>	788
741	<a href="#">An automatic metric for mt evaluation with im-</a>	Wenliang Dai, Junnan Li, Dongxu Li, Anthony	789
742	<a href="#">proved correlation with human judgments.</a> In	Meng Huat Tiong, Junqi Zhao, Weisheng Wang,	790
743	<i>IEEvaluation@ACL.</i>	Boyang Albert Li, Pascale Fung, and Steven	791
744	Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. <a href="#">Scib-</a>	C. H. Hoi. 2023. <a href="#">Instructblip: Towards general-</a>	792
745	<a href="#">ert: A pretrained language model for scientific</a>	<a href="#">purpose vision-language models with instruction</a>	793
746	<a href="#">text.</a> In <i>Conference on Empirical Methods in</i>	<a href="#">tuning.</a> <i>ArXiv, abs/2305.06500.</i>	794
747	<i>Natural Language Processing.</i>	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and	795
748	Andrés M Bran, Sam Cox, Oliver Schilter, Carlo	Kristina Toutanova. 2019. <a href="#">Bert: Pre-training</a>	796
749	Baldassari, Andrew D. White, and Philippe	<a href="#">of deep bidirectional transformers for language</a>	797
750	Schwaller. 2023. <a href="#">Chemcrow: Augmenting large-</a>	<a href="#">understanding.</a> In <i>North American Chapter of</i>	798
751	<a href="#">language models with chemistry tools.</a>	<i>the Association for Computational Linguistics.</i>	799
752	Tom B. Brown, Benjamin Mann, Nick Ryder,	Joseph L. Durant, Burton A. Leland, Douglas R.	800
753	Melanie Subbiah, Jared Kaplan, Prafulla Dhari-	Henry, and James G. Nourse. 2002. <a href="#">Reoptimiza-</a>	801
754	wal, Arvind Neelakantan, Pranav Shyam, Girish	<a href="#">tion of mdl keys for use in drug discovery.</a> <i>Jour-</i>	802
755	Sastry, Amanda Askeff, Sandhini Agarwal, Ariel	<i>nal of chemical information and computer sci-</i>	803
756	Herbert-Voss, Gretchen Krueger, T. J. Henighan,	<i>ences,</i> 42 6:1273–80.	804
757	Rewon Child, Aditya Ramesh, Daniel M.	Carl N. Edwards, T. Lai, Kevin Ros, Garrett	805
758	Ziegler, Jeff Wu, Clemens Winter, Christopher	Honke, and Heng Ji. 2022. <a href="#">Translation be-</a>	806
759	Hesse, Mark Chen, Eric Sigler, Mateusz Litwin,	<a href="#">tween molecules and natural language.</a> <i>ArXiv,</i>	807
760	Scott Gray, Benjamin Chess, Jack Clark, Christo-	<i>abs/2204.11817.</i>	808
761	pher Berner, Sam McCandlish, Alec Radford,	Carl N. Edwards, Chengxiang Zhai, and Heng Ji.	809
762	Ilya Sutskever, and Dario Amodei. 2020. <a href="#">Lan-</a>	2021. <a href="#">Text2mol: Cross-modal molecule retrieval</a>	810
763	<a href="#">guage models are few-shot learners.</a> <i>ArXiv,</i>	<a href="#">with natural language queries.</a> In <i>Conference on</i>	811
764	<i>abs/2005.14165.</i>	<i>Empirical Methods in Natural Language Pro-</i>	812
765	Feilong Chen, Minglun Han, Haozhi Zhao,	<i>cessing.</i>	813
766	Qingyang Zhang, Jing Shi, Shuang Xu, and	Li et.al. 2023. <a href="#">Blip-2: Bootstrapping language-</a>	814
767	Bo Xu. 2023. <a href="#">X-llm: Bootstrapping ad-</a>	<a href="#">image pre-training with frozen image encoders</a>	815
768	<a href="#">vanced large language models by treating</a>	<a href="#">and large language models.</a> In <i>ICML.</i>	816
769	<a href="#">multi-modalities as foreign languages.</a> <i>ArXiv,</i>	Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kang-	817
770	<i>abs/2305.04160.</i>	wei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan,	818
771	Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng,	and Huajun Chen. 2023. <a href="#">Mol-instructions: A</a>	819
772	Zhanghao Wu, Hao Zhang, Lianmin Zheng,		
773	Siyuan Zhuang, Yonghao Zhuang, Joseph E.		
774	Gonzalez, Ion Stoica, and Eric P. Xing. 2023.		

820	large-scale biomolecular instruction dataset for	100% robust molecular string representation.	866
821	large language models. <i>ArXiv</i> , abs/2306.08018.	<i>Machine Learning: Science and Technology</i> , 1.	867
822	Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lu-	Jinhyuk Lee, Wonjin Yoon, Sungdong Kim,	868
823	cas, Naoto Usuyama, Xiaodong Liu, Tristan	Donghyeon Kim, Sunkyu Kim, Chan Ho So,	869
824	Naumann, Jianfeng Gao, and Hoifung Poon.	and Jaewoo Kang. 2019. Biobert: a pre-trained	870
825	2020. Domain-specific language model pretrain-	biomedical language representation model for	871
826	ing for biomedical natural language processing.	biomedical text mining. <i>Bioinformatics</i> , 36:1234	872
827	<i>ACM Transactions on Computing for Healthcare</i>	– 1240.	873
828	( <i>HEALTH</i> ), 3:1 – 23.		
829	J. Edward Hu, Yelong Shen, Phillip Wallis,	Chunyu Li, Cliff Wong, Sheng Zhang, Naoto	874
830	Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang,	Usuyama, Haotian Liu, Jianwei Yang, Tristan	875
831	and Weizhu Chen. 2021. Lora: Low-rank	Naumann, Hoifung Poon, and Jianfeng Gao.	876
832	adaptation of large language models. <i>ArXiv</i> ,	2023a. Llava-med: Training a large language-	877
833	abs/2106.09685.	and-vision assistant for biomedicine in one day.	878
		<i>ArXiv</i> , abs/2306.00890.	879
834	Weihua Hu, Bowen Liu, Joseph Gomes, Marinka	Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao Wei, Hui	880
835	Zitnik, Percy Liang, Vijay S. Pande, and Jure	Liu, Jiliang Tang, and Qing Li. 2023b. Empow-	881
836	Leskovec. 2019. Strategies for pre-training graph	ering molecule discovery for molecule-caption	882
837	neural networks. <i>arXiv: Learning</i> .	translation with large language models: A chat-	883
		gpt perspective. <i>ArXiv</i> , abs/2306.06615.	884
838	Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and	Youwei Liang, Ruiyi Zhang, Li Zhang, and Peng	885
839	Esben Jannik Bjerrum. 2021. Chemformer: a	Xie. 2023. Drugchat: Towards enabling chatgpt-	886
840	pre-trained transformer for computational chem-	like capabilities on drug molecule graphs. <i>ArXiv</i> ,	887
841	istry. <i>Machine Learning: Science and Technol-</i>	abs/2309.03907.	888
842	<i>ogy</i> , 3.		
843	Nikhil Kandpal, Haikang Deng, Adam Roberts,	Chin-Yew Lin. 2004. Rouge: A package for au-	889
844	Eric Wallace, and Colin Raffel. 2023. Large lan-	tomatic evaluation of summaries. In <i>Annual</i>	890
845	guage models struggle to learn long-tail knowl-	<i>Meeting of the Association for Computational</i>	891
846	edge. In <i>International Conference on Machine</i>	<i>Linguistics</i> .	892
847	<i>Learning</i> , pages 15696–15707. PMLR.		
848	Jin Kim, Sera Park, Dongbo Min, and Wankyu	Haotian Liu, Chunyu Li, Yuheng Li, and	893
849	Kim. 2021. Comprehensive survey of recent	Yong Jae Lee. 2023a. Improved base-	894
850	drug discovery using deep learning. <i>Internat-</i>	lines with visual instruction tuning. <i>ArXiv</i> ,	895
851	<i>ional Journal of Molecular Sciences</i> , 22.	abs/2310.03744.	896
852	Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gin-	Haotian Liu, Chunyu Li, Qingyang Wu, and	897
853	dulyte, Jia He, Siqian He, Qingliang Li, Ben-	Yong Jae Lee. 2023b. Visual instruction tuning.	898
854	jamin A. Shoemaker, Paul A. Thiessen, Bo Yu,	<i>ArXiv</i> , abs/2304.08485.	899
855	Leonid Y. Zaslavsky, Jian Zhang, and Evan E.	Jiawei Liu, Cheng Yang, Zhiyuan Lu, Junze Chen,	900
856	Bolton. 2022. Pubchem 2023 update. <i>Nucleic</i>	Yibo Li, Mengmei Zhang, Ting Bai, Yuan Fang,	901
857	<i>acids research</i> .	Lichao Sun, Philip S. Yu, and Chuan Shi. 2023c.	902
		Towards graph foundation models: A survey and	903
858	Sunghwan Kim, Paul A. Thiessen, Tiejun Cheng,	beyond. <i>ArXiv</i> , abs/2310.11829.	904
859	Jian Zhang, Asta Gindulyte, and Evan E. Bolton.		
860	2019. Pug-view: programmatic access to chemi-	Peng Liu, Yiming Ren, and Zhixiang Ren. 2023d.	905
861	cal annotations integrated in pubchem. <i>Journal</i>	Git-mol: A multi-modal large language model	906
862	<i>of Cheminformatics</i> , 11.	for molecular science with graph, image, and	907
		text. <i>ArXiv</i> , abs/2308.06911.	908
863	Mario Krenn, Florian Hase, AkshatKumar Nigam,	Shengchao Liu, Weili Nie, Chengpeng Wang,	909
864	Pascal Friederich, and Alán Aspuru-Guzik. 2019.	Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian	910
865	Self-referencing embedded strings (selfies): A		



911	Tang, Chaowei Xiao, and Anima Anandkumar. 2022. <a href="#">Multi-modal molecule structure-text model for text-based retrieval and editing</a> . <i>ArXiv</i> , abs/2212.10789.	956
912		957
913		958
914		959
915	Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2021. <a href="#">Pre-training molecular graph representation with 3d geometry</a> . <i>ArXiv</i> , abs/2110.07728.	960
916		961
917		962
918		963
919	Zequn Liu, W. Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Yang Zhang, and Tie-Yan Liu. 2023e. <a href="#">Molxpt: Wrapping molecules with text for generative pre-training</a> . <i>ArXiv</i> , abs/2305.10688.	964
920		965
921		966
922		967
923		968
924	Zhiyuan Liu, Sihang Li, Yancheng Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023f. <a href="#">Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter</a> . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	969
925		970
926		971
927		972
928		973
929		974
930	Jieyu Lu and Yingkai Zhang. 2022a. <a href="#">Unified deep learning model for multitask reaction predictions with explanation</a> . <i>Journal of chemical information and modeling</i> .	975
931		976
932		977
933		978
934	Jieyu Lu and Yingkai Zhang. 2022b. <a href="#">Unified deep learning model for multitask reaction predictions with explanation</a> . <i>Journal of chemical information and modeling</i> .	979
935		980
936		981
937		982
938	Gen Luo, Yiyi Zhou, Tianhe Ren, Shen Chen, Xiaoshuai Sun, and Rongrong Ji. 2023a. <a href="#">Cheap and quick: Efficient vision-language instruction tuning for large language models</a> . <i>ArXiv</i> , abs/2305.15023.	983
939		984
940		985
941		986
942		987
943	Yi Luo, Kai Yang, Massimo Hong, Xingyi Liu, and Zaiqing Nie. 2023b. <a href="#">Molfm: A multi-modal molecular foundation model</a> . <i>ArXiv</i> , abs/2307.09484.	988
944		989
945		990
946		991
947	Yi Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023c. <a href="#">Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine</a> . <i>ArXiv</i> , abs/2308.09442.	992
948		993
949		994
950		995
951		996
952	OpenAI. 2023a. <a href="#">"chatgpt: A language model for conversational ai</a> .	997
953		998
954	OpenAI. 2023b. <a href="#">Gpt-4 technical report</a> . <i>ArXiv</i> , abs/2303.08774.	999
955		1000
		1001
		1002
		1003
	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. <a href="#">Training language models to follow instructions with human feedback</a> . <i>ArXiv</i> , abs/2203.02155.	
	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. <a href="#">Bleu: a method for automatic evaluation of machine translation</a> . In <i>Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02</i> .	
	Renjie Pi, Jiahui Gao, Shizhe Diao, Rui Pan, Hanze Dong, Jipeng Zhang, Lewei Yao, Jianhua Han, Hang Xu, and Lingpeng Kong Tong Zhang. 2023. <a href="#">Detgpt: Detect what you need via reasoning</a> . <i>ArXiv</i> , 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. <a href="#">Learning transferable visual models from natural language supervision</a> . In <i>International Conference on Machine Learning</i> .	
	Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. <a href="#">Exploring the limits of transfer learning with a unified text-to-text transformer</a> . <i>J. Mach. Learn. Res.</i> , 21:140:1–140:67.	
	Raghunathan Ramakrishnan, Pavlo O. Dral, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. 2014a. <a href="#">Quantum chemistry structures and properties of 134 kilo molecules</a> . <i>Scientific Data</i> , 1.	
	Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. 2014b. <a href="#">Quantum chemistry structures and properties of 134 kilo molecules</a> . <i>Scientific data</i> , 1(1):1–7.	
	B. Ramsundar, P. Eastman, P. Walters, and V. Pande. 2019. <a href="#">Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More</a> . O'Reilly.	



1004	Ahmet Sureyya Rifaioglu, Heval Atas, Maria Jesus	and Robert Stojnic. 2022. <a href="#">Galactica: A</a>	1051
1005	Martin, Rengul Cetin-Atalay, Volkan Atalay, and	<a href="#">large language model for science</a> . <i>ArXiv</i> ,	1052
1006	Tunca Dogan. 2018. <a href="#">Recent applications of deep</a>	<a href="#">abs/2211.09085</a> .	1053
1007	<a href="#">learning and machine intelligence on in silico</a>		
1008	<a href="#">drug discovery: methods, tools and databases</a> .	Hugo Touvron, Thibaut Lavril, Gautier Izacard,	1054
1009	<i>Briefings in Bioinformatics</i> , 20:1878 – 1912.	Xavier Martinet, Marie-Anne Lachaux, Timo-	1055
1010	Nadine Schneider, Roger A. Sayle, and Gregory A.	thée Lacroix, Baptiste Rozière, Naman Goyal,	1056
1011	Landrum. 2015. <a href="#">Get your atoms in order - an</a>	Eric Hambro, Faisal Azhar, Aurelien Rodriguez,	1057
1012	<a href="#">open-source implementation of a novel and ro-</a>	Armand Joulin, Edouard Grave, and Guillaume	1058
1013	<a href="#">bust molecular canonicalization algorithm</a> . <i>Jour-</i>	Lample. 2023a. <a href="#">Llama: Open and efficient founda-</a>	1059
1014	<i>nal of chemical information and modeling</i> , 55	<a href="#">tion language models</a> . <i>ArXiv</i> , <a href="#">abs/2302.13971</a> .	1060
1015	10:2111–20.		
1016	Philippe Schwaller, Teodoro Laino, Théophile	Hugo Touvron, Louis Martin, Kevin R. Stone, Pe-	1061
1017	Gaudin, Peter Bolgar, Constantine Bekas, and	ter Albert, Amjad Almahairi, Yasmine Babaei,	1062
1018	Alpha Albert Lee. 2018. <a href="#">Molecular transformer:</a>	Nikolay Bashlykov, Soumya Batra, Prajjwal	1063
1019	<a href="#">A model for uncertainty-calibrated chemical re-</a>	Bhargava, Shruti Bhosale, Daniel M. Bikel,	1064
1020	<a href="#">action prediction</a> . <i>ACS Central Science</i> , 5:1572 –	Lukas Blecher, Cristian Cantón Ferrer, Moya	1065
1021	1583.	Chen, Guillem Cucurull, David Esiobu, Jude	1066
1022	Hayal Bulbul Sonmez, Figen Kuloğlu,	Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	1067
1023	Koksal Karadag, and Fred Wudl. 2012.	Cynthia Gao, Vedanuj Goswami, Naman Goyal,	1068
1024	<a href="#">Terephthalaldehyde- and isophthalaldehyde-</a>	Anthony S. Hartshorn, Saghar Hosseini, Rui	1069
1025	<a href="#">based polyspiroacetals</a> . <i>Polymer Journal</i> ,	Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez,	1070
1026	44:217–223.	Madian Khabza, Isabel M. Kloumann, A. V. Ko-	1071
1027	Hannes Stärk, D. Beaini, Gabriele Corso, Pruden-	renev, Punit Singh Koura, Marie-Anne Lachaux,	1072
1028	cio Tossou, Christian Dallago, Stephan Gunne-	Thibaut Lavril, Jenya Lee, Diana Liskovich,	1073
1029	mann, and Pietro Lio’. 2021. <a href="#">3d infomax im-</a>	Yinghai Lu, Yuning Mao, Xavier Martinet, Todor	1074
1030	<a href="#">proves gnns for molecular property prediction</a> .	Mihaylov, Pushkar Mishra, Igor Molybog, Yixin	1075
1031	In <i>International Conference on Machine Learn-</i>	Nie, Andrew Poulton, Jeremy Reizenstein, Rashi	1076
1032	<i>ing</i> .	Rungta, Kalyan Saladi, Alan Schelten, Ruan	1077
1033	Bing Su, Dazhao Du, Zhao-Qing Yang, Yujie Zhou,	Silva, Eric Michael Smith, R. Subramanian, Xia	1078
1034	Jiangmeng Li, Anyi Rao, Haoran Sun, Zhiwu Lu,	Tan, Binh Tang, Ross Taylor, Adina Williams,	1079
1035	and Ji rong Wen. 2022. <a href="#">A molecular multimodal</a>	Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan	1080
1036	<a href="#">foundation model associating molecule graphs</a>	Zarov, Yuchen Zhang, Angela Fan, Melanie	1081
1037	<a href="#">with natural language</a> . <i>ArXiv</i> , <a href="#">abs/2209.05481</a> .	Kambadur, Sharan Narang, Aurelien Rodriguez,	1082
1038	Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin	Robert Stojnic, Sergey Edunov, and Thomas	1083
1039	Su, Suqi Cheng, Dawei Yin, and Chao Huang.	Scialom. 2023b. <a href="#">Llama 2: Open foundation and</a>	1084
1040	2023. <a href="#">Graphgpt: Graph instruction tuning for</a>	<a href="#">fine-tuned chat models</a> . <i>ArXiv</i> , <a href="#">abs/2307.09288</a> .	1085
1041	<a href="#">large language models</a> . <i>ArXiv</i> , <a href="#">abs/2310.13023</a> .		
1042	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann	Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao	1086
1043	Dubois, Xuechen Li, Carlos Guestrin, Percy	Sun, and Junzhou Huang. 2019. <a href="#">Smiles-bert:</a>	1087
1044	Liang, and Tatsunori B. Hashimoto. 2023. Stan-	<a href="#">Large scale unsupervised pre-training for molec-</a>	1088
1045	ford alpaca: An instruction-following llama	<a href="#">ular property prediction</a> . <i>Proceedings of the 10th</i>	1089
1046	model. <a href="https://github.com/tatsu-lab/stanford_alpaca">https://github.com/tatsu-lab/</a>	<i>ACM International Conference on Bioinforma-</i>	1090
1047	<a href="#">stanford_alpaca</a> .	<i>tics, Computational Biology and Health Infor-</i>	1091
1048	Ross Taylor, Marcin Kardas, Guillem Cucu-	<i>matics</i> .	1092
1049	rull, Thomas Scialom, Anthony S. Hartshorn,	Wen Wang, Zhe Chen, Xiaokang Chen, Jiannan	1093
1050	Elvis Saravia, Andrew Poulton, Viktor Kerkez,	Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong	1094
		Lu, Jie Zhou, Y. Qiao, and Jifeng Dai. 2023.	1095
		<a href="#">Visionllm: Large language model is also an open-</a>	1096
		<a href="#">ended decoder for vision-centric tasks</a> . <i>ArXiv</i> ,	1097
		<a href="#">abs/2305.11175</a> .	1098

1099	Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2021. <a href="#">Molecular contrastive learning of representations via graph neural networks</a> . <i>Nature Machine Intelligence</i> , 4:279 – 287.	1144
1100		1145
1101		1146
1102		1147
1103		1148
1104	Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. <a href="#">Finetuned language models are zero-shot learners</a> . <i>ArXiv</i> , abs/2109.01652.	1149
1105		1150
1106		1151
1107		1152
1108		1153
1109	Jinmao Wei, Xiao-Jie Yuan, Qinghua Hu, and Shuqin Wang. 2010. <a href="#">A novel measure for evaluating classifiers</a> . <i>Expert Syst. Appl.</i> , 37:3799–3809.	1154
1110		1155
1111		1156
1112		1157
1113	David Weininger. 1988. <a href="#">Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules</a> . <i>J. Chem. Inf. Comput. Sci.</i> , 28:31–36.	1158
1114		1159
1115		1160
1116		1161
1117	Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. <a href="#">Pmc-llama: Towards building open-source language models for medicine</a> .	1162
1118		1163
1119		1164
1120		1165
1121	Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay S. Pande. 2017. <a href="#">Moleculenet: A benchmark for molecular machine learning</a> . <i>arXiv: Learning</i> .	1166
1122		1167
1123		1168
1124		1169
1125		1170
1126	Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. <a href="#">Baize: An open-source chat model with parameter-efficient tuning on self-chat data</a> . <i>ArXiv</i> , abs/2304.01196.	1171
1127		1172
1128		1173
1129		1174
1130	Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. <a href="#">How powerful are graph neural networks?</a> <i>ArXiv</i> , abs/1810.00826.	1175
1131		1176
1132		1177
1133	Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. <a href="#">Gpt4tools: Teaching large language model to use tools via self-instruction</a> . <i>ArXiv</i> , abs/2305.18752.	1178
1134		1179
1135		1180
1136		1181
1137	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. <a href="#">Retroformer: Pushing the limits of interpretable end-to-end retrosynthesis transformer</a> . In <i>ICML</i> .	1182
1138		1183
1139		1184
1140		1185
1141		1186
1142		
1143		
	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. <a href="#">mplug-owl: Modularization empowers large language models with multimodality</a> . <i>ArXiv</i> , abs/2304.14178.	
	Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. <a href="#">A survey on multimodal large language models</a> . <i>ArXiv</i> , abs/2306.13549.	
	Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. <a href="#">Graph contrastive learning with augmentations</a> . <i>ArXiv</i> , abs/2010.13902.	
	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. <a href="#">Glm-130b: An open bilingual pre-trained model</a> . <i>ArXiv</i> , abs/2210.02414.	
	Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022b. <a href="#">A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals</a> . <i>Nature Communications</i> , 13.	
	Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. <a href="#">Pmc-vqa: Visual instruction tuning for medical visual question answering</a> . <i>ArXiv</i> , abs/2305.10415.	
	Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. <a href="#">Uni-mol: A universal 3d molecular representation learning framework</a> . In <i>International Conference on Learning Representations</i> .	
	Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. <a href="#">Minigt-4: Enhancing vision-language understanding with advanced large language models</a> . <i>ArXiv</i> , 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 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 corresponding instruction datasets. In InstructMol training, we

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..

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 RD-Kit 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↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	METEOR↑
<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%)
Freeze LLM in the second stage	~ 0	~ 0	~ 0	~ 0	~ 0	~ 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


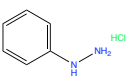
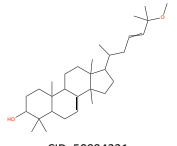
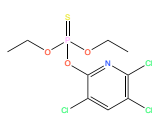
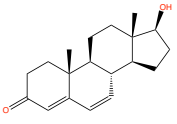
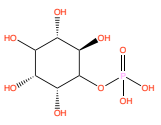
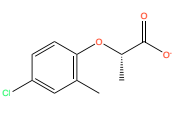
	Mol-Instruction	InstructMol	Ground Truth
 K <sup>+</sup> CID: 2735122	The molecule is the potassium salt of acetic acid. It has a role as a buffer. It is a potassium salt and an acetate salt. It contains an acetate.	The molecule is a potassium salt that is the monokis salt of carbonic acid. It has a 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 (13alpha,14beta,17alpha,20S,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 Cornus walteri. 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 (13alpha,14beta,17alpha,20S,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 Cornus walteri. 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 Homo sapiens with data available.	The molecule is a 17beta-hydroxy steroid that is testosterone in which the 17beta hydrogen is replaced by a methyl group. It has a role as an androgen. It is a 3-oxo-Delta(4) steroid and a 17beta-hydroxy steroid. It derives from a testosterone.	The molecule is a 17beta-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 17beta-hydroxy steroid and an enone. It derives from a testosterone.
 CID: 107737	The molecule is a metabolite found in or produced by Escherichia coli (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 Escherichia coli 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. It is an enantiomer of a (S)-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

1321

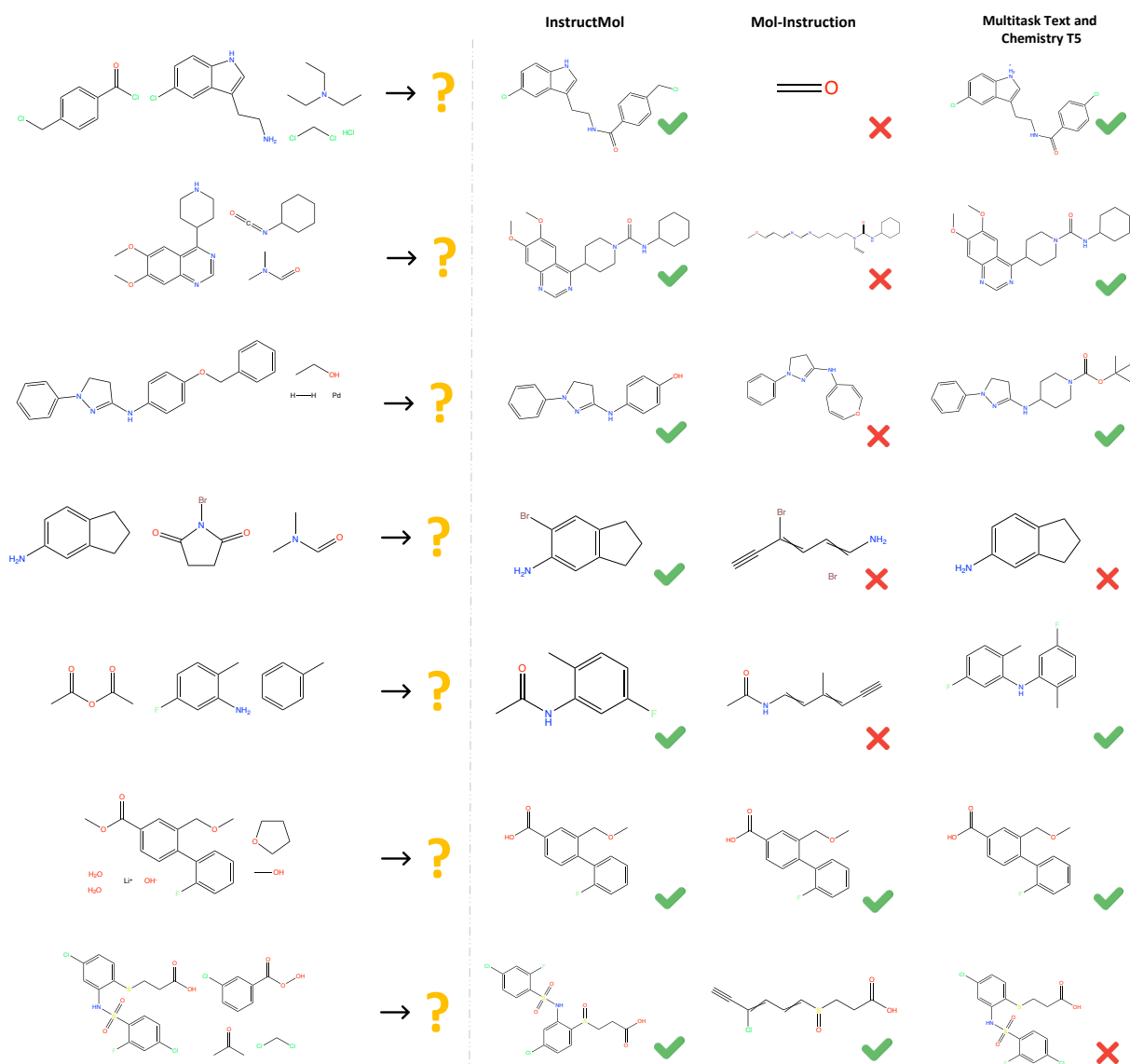


Figure 5: More examples of forward reaction prediction task. We include Mol-Instruction (Fang et al., 2023) and Multitask-Text-and-Chemistry-T5 (Christofidellis et al., 2023) as baselines.

## D.4 More Results of Reagent Prediction

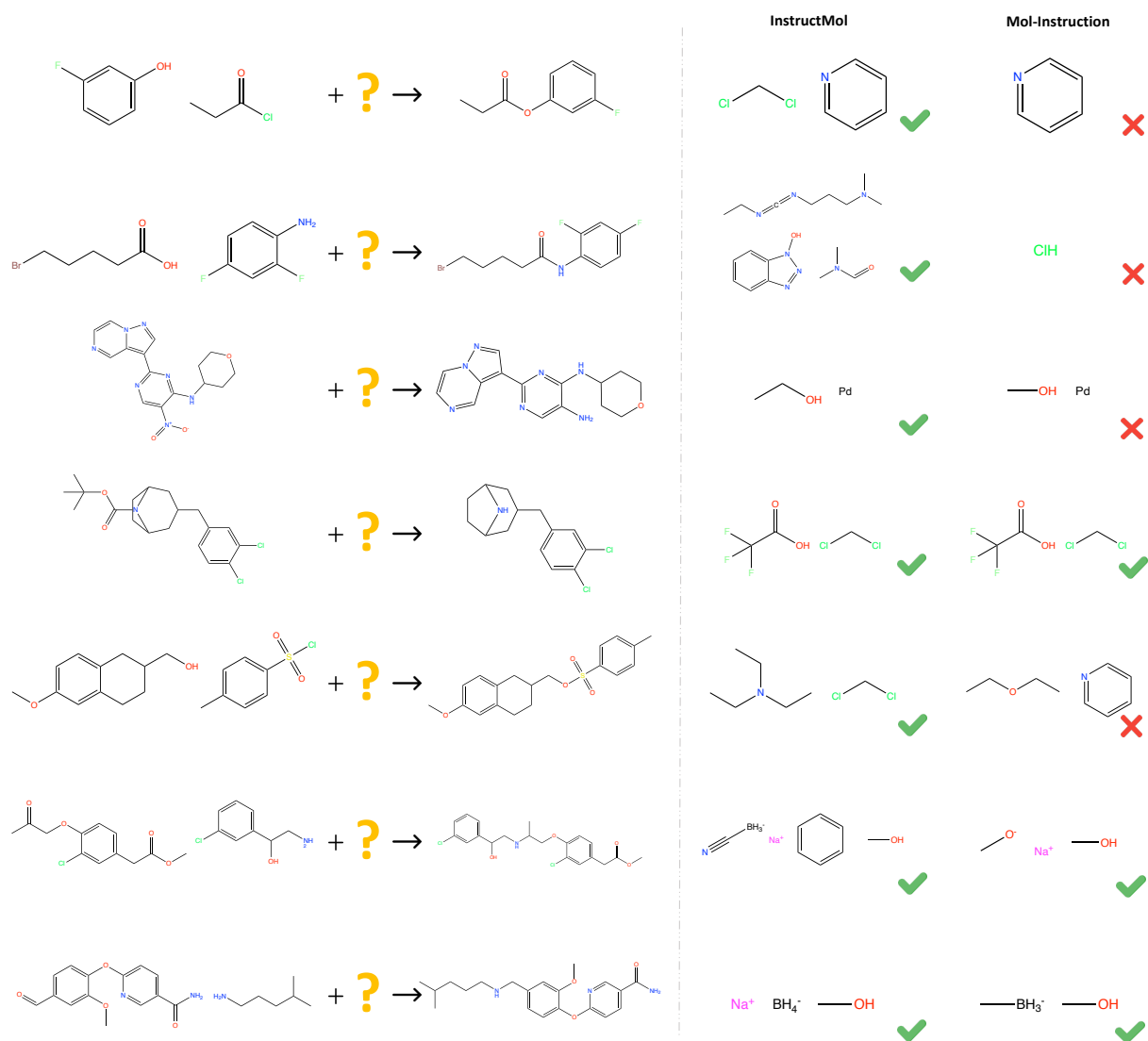


Figure 6: More examples of the reagent prediction task. We include Mol-Instruction (Fang et al., 2023) as the baseline.



## D.5 More Results of Retrosynthesis Prediction

1323

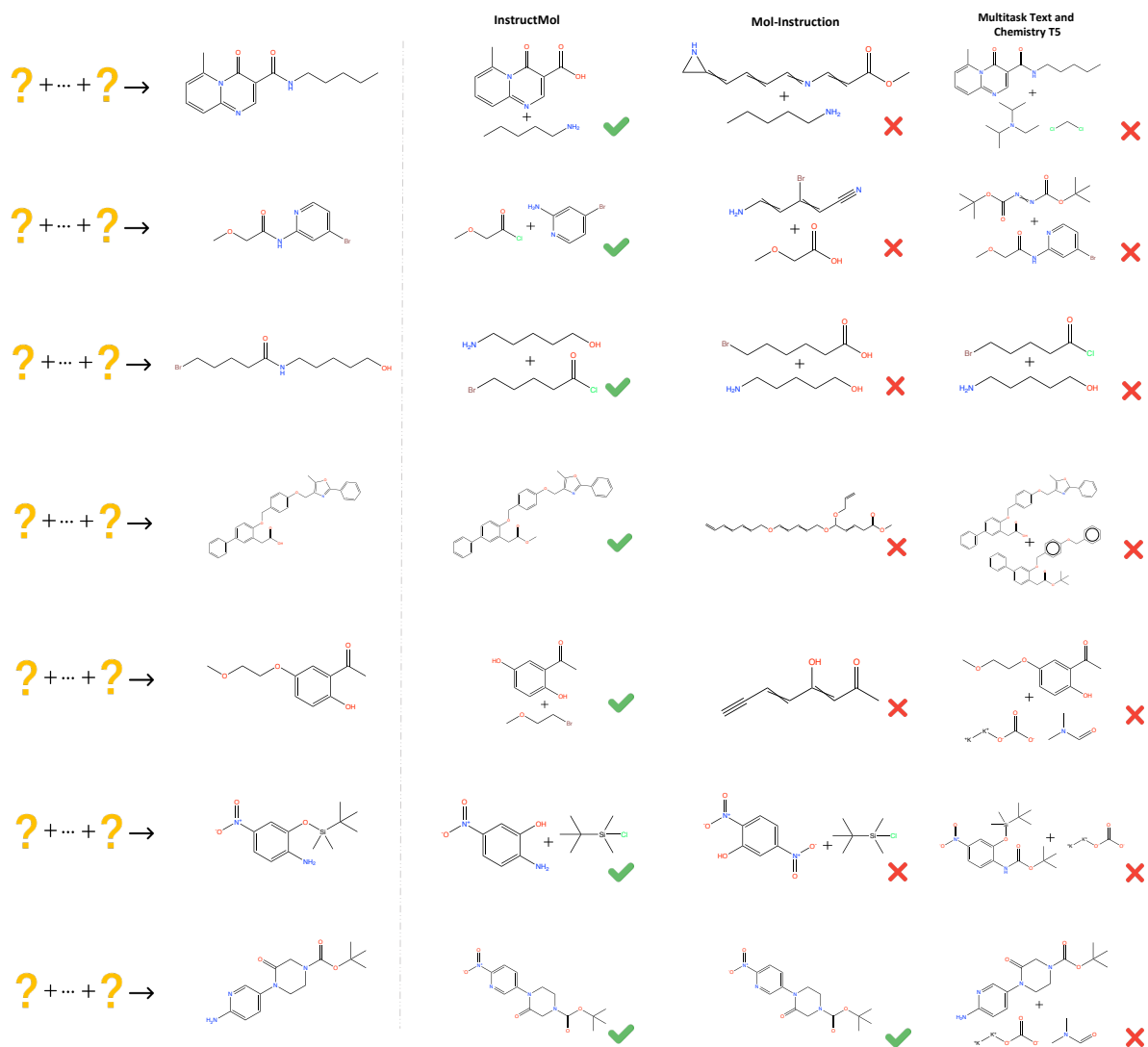


Figure 7: More examples of the retrosynthesis prediction task. We include Mol-Instruction (Fang et al., 2023) and Multitask-Text-and-Chemistry-T5 (Christofidellis et al., 2023) as baselines.

1324  
1325  
1326  
1327  
1328  
1329  
1330

## D.6 Difficult Cases

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.

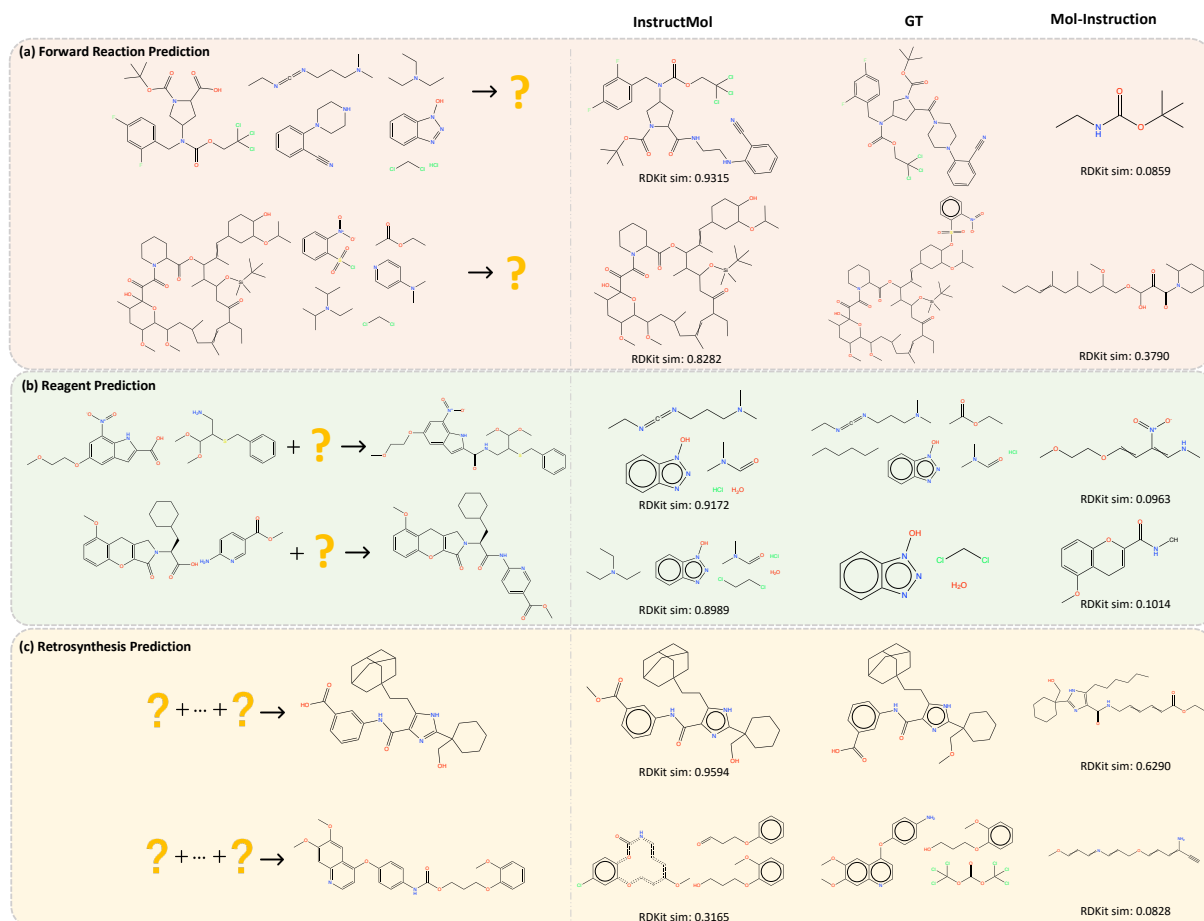


Figure 8: We present several cases with a certain degree of misalignment compared to the ground truth, accompanied by RDKit fingerprint similarity results relative to the ground truth. Due to the heightened complexity of compounds involved in chemical reactions, the difficulty of the task increases,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

1331  
1332  
1333  
1334  
1335  
1336  
1337  
1338  
1339  
1340  
1341  
1342

## 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 11.

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.

1343  
1344  
1345  
1346

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>			
CCC(=O)Cl + OC1=CC=CC(F)=C1 + ClCC1 + C2=CC=NC=C2 →?	CCC(=O)OC1=CC=CC(F)=C1	✓	✓
<i>Retrosynthesis Prediction</i>			
? → C(CCNC(=O)CCCCBr)CCO	NCCCCCO.O=C(O)CCCCBr	"Similar to controlled chemistry, reject to answer"	✓

Table 11: 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.

1347  
1348  
1349  
1350  
1351  
1352