PRESTO: Progressive Pretraining Enhances Synthetic Chemistry Outcomes

He Cao[1,2*†] **Yanjun Shao**[3*] **Zhiyuan Liu**[4] **Zijing Liu**[1]
Xiangru Tang[3] **Yuan Yao**[2] **Yu Li**[1‡]
[1]International Digital Economy Academy (IDEA)
[2]Hong Kong University of Science and Technology
[3]Yale University [4]National University of Singapore
hcaoaf@connect.ust.hk {yanjun.shao, xiangru.tang}@yale.edu
acharkq@gmail.com yuany@ust.hk {liuzijing, liyu}@idea.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) have seen growing adoption across various scientific disciplines. These advancements encourage the investigation of molecule-text modeling within synthetic chemistry, a field dedicated to designing and conducting chemical reactions to synthesize new compounds with desired properties and applications. Current approaches, however, often neglect the critical role of multiple molecule graph interaction in understanding chemical reactions, leading to suboptimal performance in synthetic chemistry tasks. This study introduces **PRESTO** (Progressive Pretraining Enhances Synthetic Chemistry Outcomes), a new framework that bridges the molecule-text modality gap by integrating a comprehensive benchmark of pretraining strategies and dataset configurations. It progressively improves multimodal LLMs through cross-modal alignment and multi-graph understanding. Our extensive experiments demonstrate that PRESTO offers competitive results in downstream synthetic chemistry tasks. The code can be found at https://github.com/IDEA-XL/PRESTO.

1 Introduction

Multi-modal Large Language Models (MLLMs) have achieved extensive success across diverse scientific domains, including medicine (Singhal et al., 2023), material science (Jablonka et al., 2023), and biochemistry (Liu et al., 2024b,a; Li et al., 2023). Motivated by these advances, molecule-text modeling emerges as a new research direction, aiming to bridge the modality gap between molecules and texts (Liu et al., 2023a; Edwards et al., 2022). These methods have shown promising results on molecule captioning, retrieval, and de-novo molecule design (Liu et al., 2024c; Edwards et al., 2021; Li et al., 2024; Tang et al., 2024a; Luo et al., 2024).

In this study, we explore molecule-text modeling within synthetic chemistry. Synthetic chemistry involves designing and executing chemical reactions to create new compounds with specific properties and applications. It is a field of immense practical value and includes tasks like forward reaction and retrosynthesis prediction. Prior molecule-text modeling works (Fang et al., 2024a; Christofidellis et al., 2023; Lu and Zhang, 2022; Zhao et al., 2024) have explored synthetic chemistry tasks, but they mostly overlook the 2D molecular graph information. However, 2D molecular graph information is crucial for understanding molecular topologies and is essential for synthetic chemistry in prior graph-based retrosynthesis studies (Somnath et al., 2021; Mao et al., 2021). On the other hand, while pioneering works (Liu et al., 2024c; Cao et al., 2023; Liu et al., 2023c; Su et al., 2022) have enabled text LLMs to perceive 2D molecular graphs, these methods struggle to process multiple 2D molecular graphs in chemical reactions, resulting in limited performance on tasks such as forward reaction prediction and reagent recommendation. This limitation stems from their inadequate exploration and analysis of multi-modal pretraining strategies (Cao et al., 2023; Luo et al., 2023c) and dataset configuration (Liang et al., 2023; Li et al., 2024), which do not fully support the comprehension of multiple graphs:

- **Multi-modal Pretraining Strategy.** The effectiveness of multi-modal LLMs is heavily influenced by their pretraining strategy (Bai et al., 2023; Lin et al., 2024; McKinzie et al., 2024), involving decisions like tuning or freezing submodules at various stages and selecting the granularity of molecular graph representations. The pretraining strategy of existing molecule-text modeling methods varies significantly (Liu et al.,

*Equal contribution.
†Work done during an internship at IDEA.
‡Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10197–10224
November 12-16, 2024

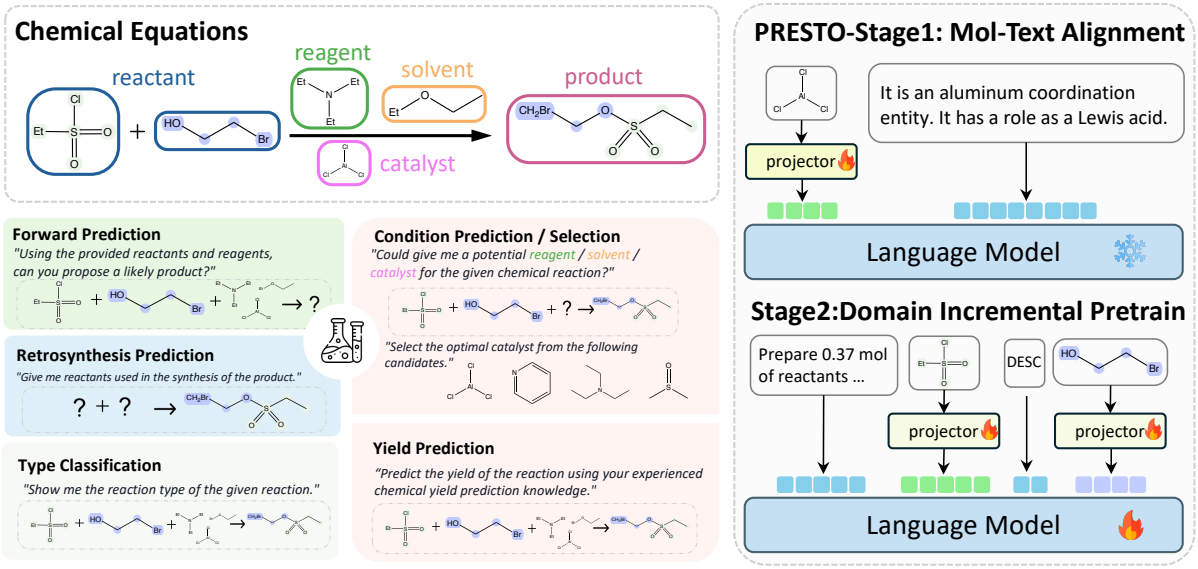


Figure 1: Panel **(top left)** illustrates the components of a prototypical chemical reaction. Panel **(bottom left)** shows the synthetic chemistry tasks that PRESTO can support as a dialogue assistant. Panel **(right)** provides an overview of the two primary stages in our Progressive Pretraining Strategy PRESTO: the Molecule-Text Alignment stage and the Domain Incremental Pretraining stage. These stages enable the evolution from single-graph text modeling to complex interleaved multi-graph text modeling.

2023c; Su et al., 2022; Liu et al., 2024c; Cao et al., 2023), creating uncertainty about the most effective approach for synthetic chemistry. Particularly, prior works notably overlook the continual pretraining on synthetic chemistry corpus, which can potentially improve performance.

- **Dataset Configuration.** The dataset plays a crucial role in the performance of LLMs. For synthetic chemistry tasks, it is evident that including data with multiple molecular graphs in context is essential. However, there is still uncertainty regarding which specific datasets (Kim et al., 2022; Lowe, 2017; Edwards et al., 2021) are most beneficial for synthetic chemistry. Additionally, it remains to be explored whether incorporating single-graph understanding tasks could further enhance performance in synthetic chemistry.

To bridge this research gap, we first present a comprehensive benchmark and the corresponding analysis for pretraining strategies and dataset configurations for synthetic chemistry. While several prior benchmarks (Fang et al., 2024a; Yu et al., 2024) overlap with synthetic chemistry, they, unfortunately, encompass a limited subset of synthetic chemistry tasks, often mishandle dataset splitting, and sometimes include potential data leakage. We prevent this by cleaning the data meticulously and generating challenging test sets with scaffold splitting. Our analysis shows that progressive multimodal domain pretraining significantly enhances reaction condition prediction accuracy. Further, we find that increasing the granularity of molecular representation and using interleaved molecule-text data with name-conversion datasets during pretraining improve downstream task performance by better leveraging domain knowledge.

Building on the insights from our benchmark, we propose Progressive Pretraining Enhances Synthetic Chemistry Outcomes (PRESTO), a specialized framework tailored for synthetic chemistry tasks. PRESTO enables a MLLM to process and understand interleaved molecular graph-text inputs, enhancing the model's understanding of the principles of chemical reaction by effectively utilizing mutual interactions between molecule-molecule and molecule-text pairs in context. To achieve this, PRESTO designs a pretraining strategy and a pretraining dataset curated for multi-graph understanding. Specifically, PRESTO improves the LLM's performance on synthetic chemistry in two stages progressively: (1) in the first training stage, PRESTO cultivates the MLLM's ability of cross-modal alignment; (2) in the second stage, PRESTO focuses on multi-graph understanding, and injects domain knowledge of synthetic chemistry into the LLM. Further, to support effective pretraining, we construct a dataset comprising ∼3 million samples of synthetic procedure descriptions and molecule name conversions. Through extensive experiments, we demonstrate that PRESTO can effectively prepare a multi-modal LLM for

downstream tasks of synthetic chemistry.

2 Related Works

Deep Learning for Synthetic Chemistry. Synthetic chemistry, a fundamental problem in chemistry, has seen significant advances through deep learning models that assist in various reaction-related tasks using descriptor-based (Segler and Waller, 2017; Segler et al., 2018), graph-based (Dai et al., 2019; Tu and Coley, 2021), and sequence-based approaches (Schwaller et al., 2019; Irwin et al., 2022). Recent works (Lu and Zhang, 2022; Schwaller et al., 2020; Fang et al., 2024a; Yu et al., 2024) also adapt language models for tasks such as forward reaction prediction (Schwaller et al., 2019), retrosynthesis (Wan et al., 2022; Liu et al., 2024d), and reaction type classification (Schwaller et al., 2021a), demonstrating high accuracy. Although these models specialize in specific synthetic chemistry tasks, their pretraining on domain-specific data limits their ability to generalize and adapt to other synthetic tasks. To address this issue, multi-task methods (Lu and Zhang, 2022; Christofidellis et al., 2023) have been explored and demonstrate strong capabilities across domains. However, they are constrained by using only molecular sequences as input, overlooking the potential of textual information to assist in modeling. In contrast, our approach integrates reaction-related textual information with molecular modeling, enabling a flexible adaptation to various downstream tasks.

Molecule & Text Modeling (MTM). The integration of biomolecular modeling with natural language leverages rich textual data sources to enhance understanding and facilitate downstream text-related molecular tasks (Edwards et al., 2022; Christofidellis et al., 2023; Pei et al., 2023; Fang et al., 2024a; Yu et al., 2024; Luo et al., 2023b). Various approaches have been proposed to learn effective representations of molecules, including 1D sequences (Fang et al., 2024b; Irwin et al., 2022; Edwards et al., 2022; Schwaller et al., 2019; Wang et al., 2019), 2D graphs (Rong et al., 2020; Ying et al., 2021; Wang et al., 2022b; Liu et al., 2023d), 3D conformations (Liu et al., 2022; Zhou et al., 2023) and a combination of them (Luo et al., 2023a; Tang et al., 2024b). Cross-modalities modeling includes contrastive learning over molecules and text (Su et al., 2022; Liu et al., 2023a; Tang et al., 2024b) or unified alignment of the two modalities through language modeling (Zeng et al., 2022; Zhao et al., 2023; Liu et al., 2023c; Li et al., 2024). Prior works have primarily focused on individual molecule understanding or molecule-text retrieval, while our research expands to model multiple molecules and contextual text, thereby facilitating tasks relevant to chemical reactions.

Multi-modal Language Models. The multi-modal large language models (MLLMs) field has seen rapid progress recently. Several works (Alayrac et al., 2022; Wang et al., 2022a; Chen et al., 2023; Dai et al., 2023; Li et al., 2023; Huang et al., 2023; Liu et al., 2024b) have proposed different architectures for integrating visual information into LLMs. Researchers have explored various strategies for integrating external modalities into LLMs. Lin et al. (2024) and McKinzie et al. (2024) conducted ablation studies on textual and visual data composition during training. Karamcheti et al. (2024) examined the design space of MLLMs, including training pipeline, modality representations, and scaling. Recent studies have attempted to apply similar methods to small molecule (Li et al., 2024; Cao et al., 2023; Liang et al., 2023) or protein domains (Wang et al., 2023b; Liu et al., 2024e). However, there are very few studies investigating the specific design of training strategies in the biomolecular domain.

3 PRESTO Framework

3.1 Preliminary

Here we introduce our model architecture, which follows the common practice in multi-modal LLMs (Liu et al., 2024b; Bai et al., 2023; Karamcheti et al., 2024). Formally, our model processes a collection of 2D molecule graphs represented as $\{\boldsymbol{X}_G^{(i)}\}_{i=1}^{n}$, along with text prompt tokens $\{\boldsymbol{X}_T^{(j)}\}_{j=1}^{m}$ describing synthetic processes or task queries. The input sequence is designed to accommodate the interleaved nature of text and molecule tokens, denoted $\{t_k\}_{k=1}^{m+n}$, where each t_k is a text token $\boldsymbol{X}_T^{(j)}$ or a molecule graph $\boldsymbol{X}_G^{(i)}$. These inputs are processed through 1) a molecular representation encoder, 2) a molecule-language projector, and 3) a language model.

Molecular Representation. Each $\boldsymbol{X}_G^{(i)}$ is first processed by a molecule encoder f_M, which outputs a sequence of features $p_M^{(i)}$, such that $p_M^{(i)} = f_M(\boldsymbol{X}_G^{(i)})$. The length of $p_M^{(i)}$ is variable and depends on the granularity of the representation.

Molecule-Language Projector. Next, we map

each $p_M^{(i)}$ to embeddings $e_M^{(i)}$ using a learned projector f_ψ, where $e_M^{(i)} = f_\psi(p_M^{(i)})$.

Language Model. The interleaved input sequence $\mathcal{E}_I$ is formed by the ordered union of molecule embeddings $\mathcal{E}_M = \{e_M^{(i)}\}_{i=1}^n$ and text token embeddings $\mathcal{E}_T = \{e_T^{(j)} | e_T^{(j)} = f_{\text{embed}}(\boldsymbol{X}_T^{(j)})\}_{j=1}^m$:

$$\mathcal{E}_I = \mathcal{E}_M \cup_o \mathcal{E}_T,$$

where $\cup_o$ preserves the order of elements as they appear in the original input sequence $\{t_k\}_{k=1}^{m+n}$. This interleaved sequence is passed to the language model to generate the output text $\boldsymbol{X}_O = \text{LM}_\theta(\mathcal{E}_I)$.

3.2 Training Procedure

Our complete training procedure includes the PRESTO's two-stage pretraining and the downstream supervised finetuning.

PRESTO-Stage1: Molecule-Text Alignment. This stage aims to bridge the modality gap between the molecular and textual representations. We start from a pretrained molecule encoder f_M, a language model LM_θ, and a randomly initialized molecule-language projector f_ψ. f_ψ is then trained on molecule-text pairs from (Kim et al., 2022) while freezing the weights of f_M and LM_θ. The template for captioning can be found in Appendix D.1.

PRESTO-Stage2: Domain Incremental Pretraining. During this stage, we continue to train the model on a large corpus of molecule-text pairs with interleaved segments (Lowe, 2017; Kim et al., 2022). Training on mixed data helps the model further understand the relationships between molecular graphs and text. Both f_M and LM_θ are updated in this stage. See Appendix D.1 for details of the instruction template.

Supervised Fine-Tuning (SFT). The final stage adapts the pretrained model to a diverse set of downstream tasks by instruction tuning. Similar to (Cao et al., 2023; Liu et al., 2023c), each example consists of input molecules or reactions $\{\boldsymbol{X}_G^{(i)}\}_{i=1}^n$, a natural language instruction $\{\boldsymbol{X}_T^{(j)}\}_{j=1}^m$, and the target output $\boldsymbol{X}_O$. Details of the instruction template can be found in the Appendix D.2.

3.3 Pretrain Dataset

We present datasets utilized in the PRESTO pretraining pipeline. For the first stage of alignment, we use a caption dataset, and for the second stage of domain incremental pretraining, we use an interleaved molecule-text and name-conversion dataset.

TASK	# TRAIN	# VALID	# TEST	# ALL
Pretrain Stage1: Molecule Caption				
DATA SOURCE: Kim et al. (2022)				
PubChem Caption	326,675	-	-	326,675
Pretrain Stage2: Interleaved Molecule-Text				
DATA SOURCE: Lowe (2017)				
USPTO-Application	1,588,709	-	-	1,588,709
Pretrain Stage2: Name Conversion				
DATA SOURCE: Kim et al. (2022); Yu et al. (2024)				
IUPAC to Formula	300,000	1,497	2,993	304,490
IUPAC to SMILES	300,000	1,497	2,993	304,490
Molecule Graph to Formula	300,000	1,497	2,993	304,490
Molecule Graph to IUPAC	300,000	1,497	2,993	304,490
Molecule Graph to SMILES	293,288	-	-	293,288

Table 1: PRESTO progressive pretraining dataset.

Caption Dataset. We use molecule-text pairs sourced from PubChem (Kim et al., 2022) for aligning molecule and text modalities. Each molecule structure is associated with a textual description of chemical and physical properties or high-level bioactivity information.

Interleaved Molecule-Text Dataset. We start by extracting raw descriptions of experimental procedures from the chemical reaction database USPTO-Applications (Lowe, 2017). Further, we use BERN2 (Sung et al., 2022) to identify all molecule entities in the texts and convert them into 2D molecular graphs. We then preprocess the data to remove samples with too many molecule entities or molecules with excessive atom counts to control input length. The resulting interleaved dataset comprises approximately 1.6M samples, covering more than 342K unique molecules. Refer to Appendix A.2 for detailed processing steps and data statistics.

Name Conversion Dataset. A molecule can be represented by 2D molecular graphs and different 1D sequential representations: IUPAC names (Favre and Powell, 2014), chemical formulas (Hill, 1900), and SMILES (Weininger, 1988). These 1D sequential representations are used interchangeably in the textual corpus, and each corresponds to a particular aspect of molecular structures. For example, the IUPAC name highlights the subgraph composition of molecules, while SMILES explicitly lists all atom types. Therefore, learning the conversion between these 1D representations and 2D graphs helps the LLM to align different molecular mentions in texts and improves its understanding of molecular structures.

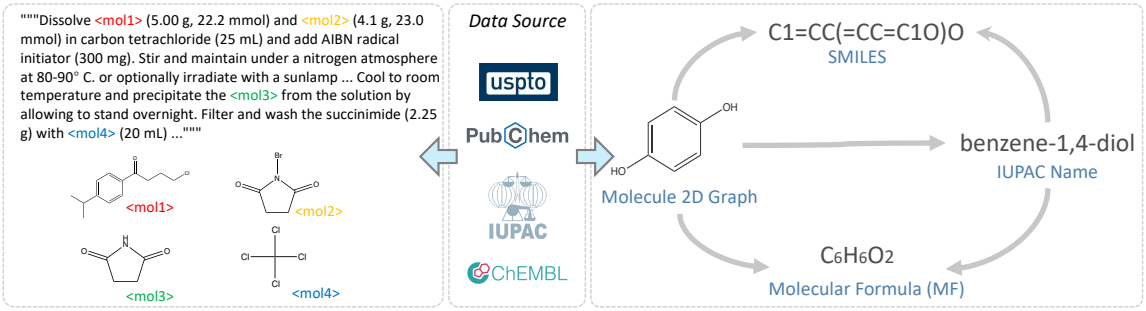


Figure 2: Panel **(a)** illustrates the interleaved molecule-text dataset format, primarily derived from USPTO-Application (Lowe, 2017). Panel **(b)** displays the five tasks included in the Molecular Name Conversion Tasks (directions drawn as arrows), with data mainly sourced from PubChem (Kim et al., 2022), IUPAC (Favre and Powell, 2014), and ChEMBL (Zdrazil et al., 2023).

3.4 Downstream Tasks

We evaluate PRESTO on a diverse set of downstream tasks in synthetic chemistry, as detailed in Table 2. Our assessment provides a more comprehensive and representative evaluation of downstream tasks, extending beyond the scope of previous benchmarks. The detailed data preprocessing pipeline is provided in the Appendix A.3.

Reaction Prediction. This category includes two tasks: *Forward Prediction*, which involves predicting the product molecules given the reactant molecules, and *Retrosynthesis*, which predicts the reactant molecules given the target product molecule. Data from USPTO-full (Lowe, 2017; Yu et al., 2024) and USPTO_500_MT (Irwin et al., 2022; Fang et al., 2024a) are used for these tasks.

Reaction Condition Prediction. This category involves predicting the *reagents*, *catalysts*, and *solvents* for a given reaction. We utilize extracted reaction condition information from Qian et al. (2023) and re-split the reagent prediction dataset provided by Fang et al. (2024a) into three separate sets.

Reagent Selection. This task, also known as reagent recommendation, involves identifying the most suitable reagents for a specific chemical reaction or process. It is divided into three categories: reactant selection, ligand selection, and solvent selection. We formulate it as choosing the most suitable reagent from a list of candidates. We adopt the dataset collected from Guo et al. (2023).

Reaction Type Classification. This task aims to classify a reaction into predefined types to navigate chemical space and better understand the underlying mechanisms. We use the USPTO 1K TPL dataset from Schwaller et al. (2021a) with 1000 labeled classes. Learned representations can also serve as reaction fingerprints, capturing fine-grained differences.

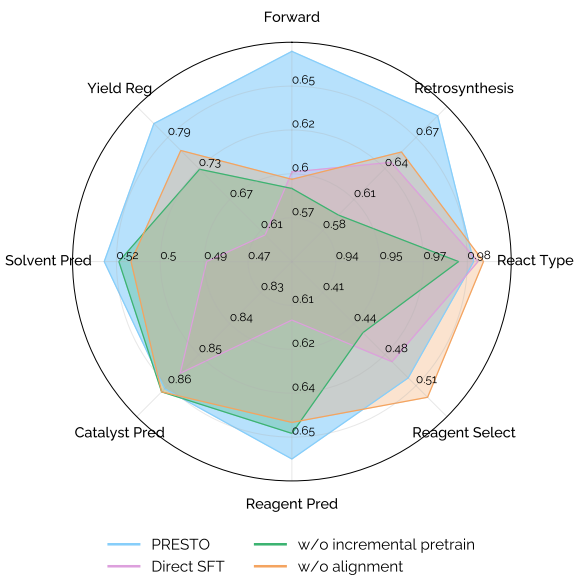
Yield Regression. This task involves estimating the amount of product (yield) obtained from a given chemical reaction. We test the model's performance on two High-Throughtput experimentation (HTE) datasets: *Buchwald-Hartwig* and *Suzuki-Miyaura*. Both datasets are obtained from Schwaller et al. (2021b).

Remark: Generating an Uncontaminated and Challenging Test Set. Data leakage is commonly observed in recent LLM studies (Blevins and Zettlemoyer, 2022; Deng et al., 2024; Li and Flanigan, 2024), and we have observed the same issue in early benchmarks of chemical reaction prediction (Fang et al., 2024a). This issue leads to skewed evaluation and can hinder the development of truly effective models. To present a reliable chemical reaction task evaluation, we meticulously ensure no overlap between our pretraining/training datasets and testing datasets. Further, we establish a test set for the reaction prediction task by including only samples with a scaffold similarity below a certain threshold compared to the training samples. This approach separates the training and testing distributions, improving the robustness and accuracy of our evaluations. Prior benchmarks often used random splits, resulting in significant overlaps in molecular scaffolds between training and test sets, compromising the evaluation of real-world generalization. For further details, please refer to the Appendix A.1.

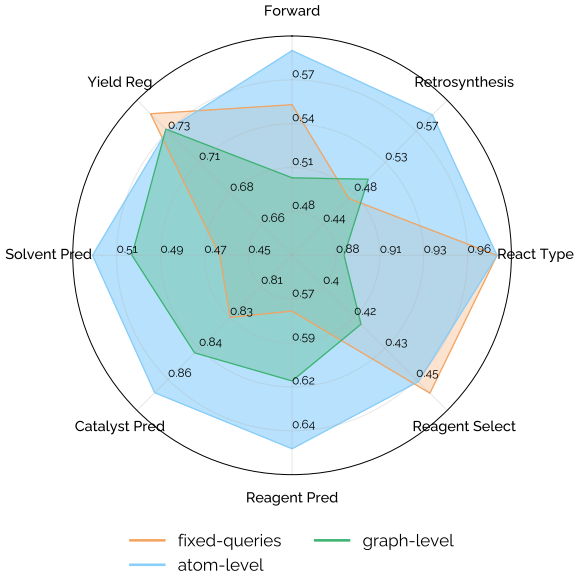
4 Analyzing Pre-Training Strategy and Dataset Configuration

In this section, we conduct experiments to evaluate the impact of different pretraining strategies and dataset configurations on downstream tasks.

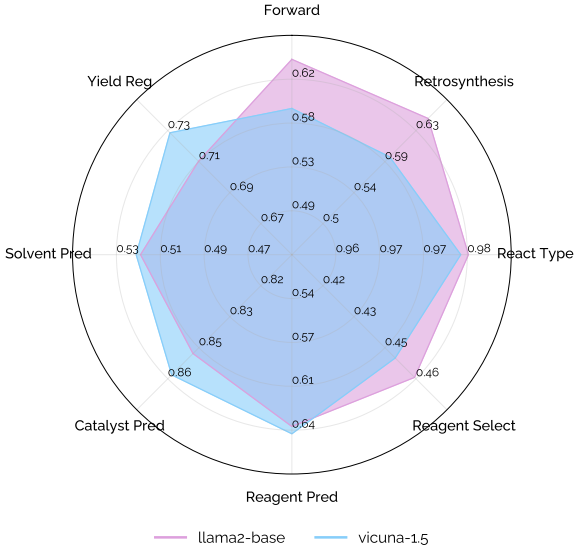
Experimental Setting. We use the GIN (Xu et al., 2019) pretrained by MoleculeSTM (Liu et al., 2023a) as the default graph encoder f_M and a two-



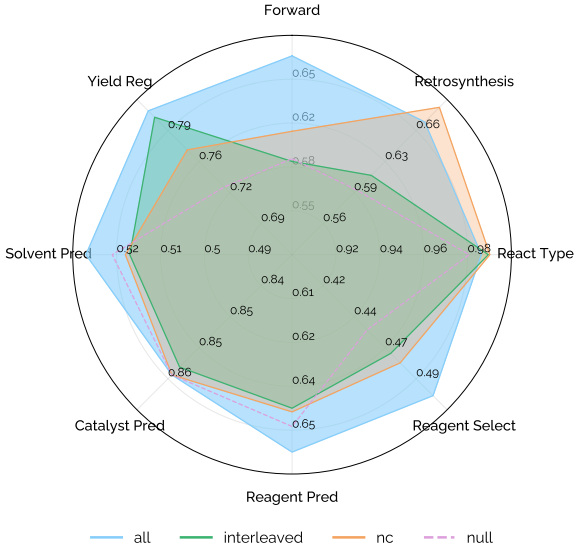
(a) **Multi-stage Pretrain Strategy Ablation.**



(b) **Molecule Token Granularity Ablation.**



(c) **Base vs. Instruct-Tuned LLMs.**



(d) **PRESTO-Stage2 Dataset Configuration Ablation.**

Figure 3: **Performance analysis of different pretraining strategies and dataset configurations. (a)** Ablation study on the multi-modal pretraining strategy. **(b)** We explore various options for the granularity of molecular encoded tokens. **(c)** Comparison between base (Llama-2) and instruct-tuned (Vicuna v1.5) language models. **(d)** Ablation study on dataset configuration for PRESTO domain incremental pretraining stage.

layer MLP as the projector f_ψ. For the base LM_θ, we use Vicuna v1.5-7B (Chiang et al., 2023) by default. We report the mean similarity measured by Morgan (Schneider et al., 2015), MACCS (Durant et al., 2002), RDKit (Landrum et al., 2024) fingerprints for generation tasks, Top-1 accuracy for classification tasks, and R^2 scores for regression tasks. Detailed experimental settings are available in Appendix B.

4.1 Analyzing Pretraining Strategy

We investigate the impact of different pretraining strategies, varying levels of molecular representation granularity, and different LLMs on the model's performance in downstream tasks. We divide the pretraining pipeline into two stages: alignment and domain incremental pretraining, as mentioned in Section 3.2. Due to the high time and computation costs of the incremental pretraining stage, we skip it unless explicitly stated otherwise.

Finding 1: Progressive pretraining strategy enhances downstream task performance. As shown in Figure 3a, Direct SFT significantly degrades the prediction of reaction conditions and yields. This degradation occurs because the model

TASK	# TRAIN	# VALID	# TEST	# ALL
Reaction Prediction				
DATA SOURCE: Lu and Zhang (2022); Yu et al. (2024); Fang et al. (2024a)				
Forward Prediction	124,384	-	1,000	125,384
Retrosynthesis Prediction	124,384	-	1,000	125,384
Reaction Condition Prediction				
DATA SOURCE: Qian et al. (2023); Guo et al. (2023); Fang et al. (2024a)				
Reagent Prediction	57,162	6,216	6,378	69,756
Catalyst Prediction	10,232	1,059	1,015	12,306
Solvent Prediction	70,988	7,694	7,793	86,475
Reaction Condition Recommendation				
DATA SOURCE: Guo et al. (2023)				
Reagent Selection	3,955	-	300	4,255
Reaction Type Classification				
DATA SOURCE: Schwaller et al. (2021a)				
Reaction Type Classification	360,379	40,059	44,511	445,115
Yield Prediction				
DATA SOURCE: Schwaller et al. (2021b)				
Buchwald-Hartwig	3,855	-	100	3,955
Suzuki-Miyaura	5,660	-	100	5,760

Table 2: PRESTO downstream tasks dataset statistics.

must simultaneously learn to align different modalities and adapt to various downstream tasks, increasing the optimization difficulty. W/o alignment demonstrates that the alignment stage, which acts as a warm-up for modality fusion, effectively connects molecular and language information, aiding the transition of a general-domain LLM to the chemistry domain. Additionally, w/o incremental pretrain underscores the importance of domain incremental pretraining in enhancing multi-graph modeling and domain knowledge adaptation.

Finding 2: Molecular representation granularity matters. Drawing from prior VLMs research (Karamcheti et al., 2024; Lin et al., 2024), enhancing visual resolution improves downstream performance by capturing intricate details. Similarly, we utilize various granularities for molecular representation, including graph-level (a global token per graph), atom-level (each atom represented by one token), and fixed-length query-encoding (Li et al., 2024; Liu et al., 2023c). In Figure 3b, scaling to the atom level yields substantial improvements across all tasks compared to graph-level modeling. Interestingly, the query-encoding approach performs remarkably well in regression and classification tasks but severely underperforms in tasks that require generating entire molecules. We speculate that the learned queries may fail to capture fine-grained molecular structures, resulting in suboptimal performance in generating full molecules.

Finding 3: Base and instruct-tuned LLMs demonstrate comparable capabilities. Instruct tuning is a method to finetune base LLMs (trained for next-token prediction) to function as dialogue agents that can follow instructions more effectively. Modern VLMs research (Liu et al., 2024b; Lin et al., 2024) often use instruct-tuned models like Vicuna as the base LLMs. We evaluate the impact of instruct-tuned LLM on downstream synthetic chemistry tasks via a head-to-head comparison between a base LLM (Llama-2-7B (Touvron et al., 2023)) and its instruct-tuned variant (Vicuna v1.5). Figure 3c shows that instruction-tuned LLMs slightly outperform base in reaction condition prediction and yield tasks, while base LLMs excel in forward prediction and retrosynthesis.

4.2 Analyzing Dataset Configuration

Here, we analyze the impact of dataset configurations on domain incremental pretraining.

Finding 4: Both interleaved data and name-conversion data play crucial roles in domain incremental pretraining. As shown in Figure 3d, relying solely on an interleaved molecule-text dataset can improve model performance in retrosynthesis, classification, and regression tasks, but the improvement is marginal. We believe this is because interleaved data lack strict molecule-text correspondence, making it difficult for the model to use the surrounding text to learn molecular syntax and semantics and recognize molecular structural patterns. Therefore, we introduce a name conversion task dataset to enhance contextual learning, which aids tasks requiring a deeper understanding of chemical entities and their functions. Experiments demonstrate that incrementally, pretraining with a blend of interleaved data and name conversion data better leverages the domain knowledge from the synthetic procedure corpus, facilitating downstream tasks.

5 Comparison with the State-of-the-arts

We integrate the above findings to inform our PRESTO framework at the 7B parameter scale. We present results comparing PRESTO with previous domain expert models (Irwin et al., 2022; Schwaller et al., 2019; Wan et al., 2022; Schwaller et al., 2021a; Wang et al., 2022c; Probst et al., 2022; Ahneman et al., 2018; Kwon et al., 2022; Schwaller et al., 2021b) and other LLM-based methods (Fang et al., 2024a; Livne et al., 2023; Christofidellis et al., 2023; Yu et al., 2024; Taylor et al., 2022; Zhao et al., 2024; Lu and Zhang, 2022).

Table 3 presents the performances for generation tasks. We report commonly used metrics in the MTM domain, including Exact Match, BLEU (Pa-

MODEL	EXACT↑	BLEU↑	LEVENSHTEIN↓	RDK FTS↑	MACCS FTS↑	MORGAN FTS↑	VALIDITY↑
Forward Reaction Prediction							
Chemformer* (Irwin et al., 2022)	0.372	0.824	8.097	0.755	0.820	0.717	0.994
MoleculeTransformers* (Schwaller et al., 2019)	0.313	0.663	11.735	0.549	0.619	0.532	0.925
Mol-Instruction (Fang et al., 2024a)	0.065	0.428	24.076	0.260	0.430	0.249	**0.999**
LLama2-7b* (Touvron et al., 2023)	0.251	0.658	13.167	0.533	0.630	0.512	0.940
Vicuna v1.5-7b* (Chiang et al., 2023)	0.250	0.659	12.506	0.513	0.600	0.495	0.903
LlaSMol-Mistral (Yu et al., 2024)	0.055	0.750	15.558	0.221	0.471	0.202	0.788
nach0-base (Livne et al., 2023)	0.331	**0.857**	13.108	0.628	0.709	0.594	0.977
Text+Chem T5 (Christofidellis et al., 2023)	0.236	0.750	13.631	0.523	0.630	0.505	0.967
T5Chem (Lu and Zhang, 2022)	0.313	0.703	13.632	0.535	0.616	0.520	0.965
PRESTO	**0.355**	0.836	**10.647**	**0.646**	**0.726**	**0.624**	0.973
Retrosynthesis Prediction							
Chemformer*	0.011	0.611	21.073	0.659	0.730	0.574	0.998
Retroformer* (Wan et al., 2022)	0.273	0.769	14.768	0.690	0.782	0.647	0.952
Mol-Instruction	0.039	0.395	31.611	0.279	0.478	0.26	**1.0**
LLama2-7b*	0.220	0.754	15.695	0.649	0.747	0.608	0.933
Vicuna v1.5-7b*	0.220	0.756	15.692	0.658	0.758	0.616	0.943
LlaSMol-Mistral	0.010	0.694	19.719	0.148	0.317	0.119	0.530
nach0-base	0.173	0.854	18.883	0.574	0.668	0.515	0.892
Text+Chem T5	0.042	0.620	13.952	0.261	0.281	0.206	0.345
T5Chem	0.208	0.725	17.278	0.595	0.662	0.566	0.994
PRESTO	**0.275**	**0.902**	**14.433**	**0.655**	**0.747**	**0.619**	0.980
Reaction Condition Prediction (Reagent)							
LLama2-7b*	0.312	0.564	9.058	0.560	0.575	0.466	**1.0**
Vicuna v1.5-7b*	0.315	0.585	8.664	0.576	0.587	0.473	**1.0**
nach0-base	0.001	0.172	34.212	0.053	0.134	0.039	0.932
Mol-Instruction	0.0	0.219	27.108	0.034	0.098	0.030	**1.0**
T5Chem	0.019	0.559	11.044	0.366	0.461	0.374	0.994
PRESTO	**0.458**	**0.776**	**6.206**	**0.678**	**0.683**	**0.601**	0.999
Reaction Condition Prediction (Catalyst)							
LLama2-7b*	0.680	0.720	2.545	0.882	0.868	0.687	**1.0**
Vicuna v1.5-7b*	0.685	0.703	2.451	0.883	0.869	0.692	**1.0**
nach0-base	0.0	0.072	36.442	0.129	0.055	0.009	0.849
Mol-Instruction	0.0	0.110	28.424	0.031	0.045	0.015	0.999
T5Chem	0.022	0.346	13.408	0.146	0.268	0.200	0.996
PRESTO	**0.768**	**0.814**	**1.755**	**0.914**	**0.895**	**0.774**	**1.0**
Reaction Condition Prediction (Solvent)							
LLama2-7b*	0.311	0.462	3.819	0.452	0.48	0.417	**1.0**
Vicuna v1.5-7b*	0.320	0.436	3.809	0.459	0.486	0.427	**1.0**
nach0-base	0.0	0.072	36.442	0.129	0.055	0.009	0.849
Mol-Instruction	0.0	0.155	25.117	0.030	0.122	0.035	**1.0**
T5Chem	0.083	0.311	16.224	0.458	0.424	0.397	0.995
PRESTO	**0.419**	**0.695**	**2.758**	**0.529**	**0.547**	**0.506**	0.912

Table 3: Comparison of various models on forward reaction prediction, retrosynthesis prediction, and reaction condition prediction tasks. Model indicates a domain expert method, and * denotes our re-implementation.

METHOD	REACTANT	SOLVENT	LIGAND
Reagent Selection			
LLama2-7b*	0.670	0.550	0.010
Vicuna v1.5-7b*	0.690	0.580	0.440
GPT-4†	0.299	0.526	**0.534**
GAL-30B† (Taylor et al., 2022)	0.107	0.104	0.030
LLama2-13b-chat†	0.145	0.050	0.284
ChemDFM-13b (Zhao et al., 2024)	0.240	0.120	0.350
PRESTO	**0.780**	**0.630**	0.520

METHOD	ACC↑	CEN↓	MCC↑
Reaction Type Classification			
BERT classifier (Schwaller et al., 2021a)	0.989	0.006	0.989
ContraGIN (Wang et al., 2022c)	0.993	0.001	0.993
DRFP (Probst et al., 2022)	0.977	0.011	0.977
T5Chem	**0.995**	**0.003**	**0.995**
LLama2-7b*	0.804	0.079	0.803
Vicuna v1.5-7b*	0.888	0.048	0.887
PRESTO	0.991	0.004	0.991

METHOD	B-H	S-M
Yield Regression		
DFT (Ahneman et al., 2018)	0.920	-
UAGNN (Kwon et al., 2022)	0.969	0.884
YieldBERT (Schwaller et al., 2021b)	0.950	0.815
T5Chem	**0.970**	-
LLama2-7b*	-0.476	0.121
Vicuna v1.5-7b*	-0.131	0.151
PRESTO	0.944	**0.652**

Table 4: Comparison with baselines on reagent selection, reaction type classification, and yield regression tasks. † denotes results from (Zhao et al., 2024). For reagent selection, we report the result in top-1 accuracy except for LIGAND SELECTION, where we report the top 50% accuracy. For yield regression, we report the R^2 score.

pineni et al., 2001), Levenshtein distance, Validity, and fingerprint similarities (RDKit, MACCS, and Morgan). Table 4 reports on regression and classification tasks, evaluating metrics such as Accuracy, Confusion Entropy of the confusion matrix (CEN), Matthews Correlation Coefficient (MCC), and R^2 scores. Results show that PRESTO outperforms the baseline LLMs across all downstream tasks and narrows the gap with domain expert models. These improvements highlight the effectiveness of our proposed progressive pretraining strategy and comprehensive analytical design. However, it is noteworthy that there is still room for improvement in validity. Future efforts could involve replacing SMILES with SELFIES (Krenn et al., 2019) to enhance robustness in representation.

6 Conclusion and Future Work

This study explores integrating multimodal LLMs into synthetic chemistry tasks to overcome the molecule-text modality gap. We highlight the importance of multi-graph datasets and progressive pretraining methods, showing significant improvements in reaction predictions and synthetic chem-

istry tasks. As a result, we introduce PRESTO, which outperforms baseline LLMs.

Meanwhile, current multimodal molecule models are limited to generating only 1D sequences. As a potential direction, we envision developing models capable of producing comprehensive molecular representations (i.e., 2D, 3D). Future research could also expand the diversity of datasets to include more molecular structures and improve the LLM's capability for dialogue. We aim to advance the fields of synthetic chemistry and compound discovery, ultimately creating a more powerful and versatile assistant for chemists.

7 Limitations

Despite the significant advancements achieved by PRESTO, several limitations remain. Firstly, we did not conduct ablation studies on additional molecular modalities, such as 3D structure information, nor did we explore whether combining different modalities could further enhance molecular representations and improve downstream performance. Secondly, we observed that the model's ability to answer general domain questions declined as domain-specific finetuning (SFT) progressed. Future training should consider integrating general domain SFT datasets to prevent the forgetting issue. Lastly, our base LLM is a general-domain model, and the fields of chemistry and molecular science lack specialized LLMs with parameter scales comparable to models like LLaMA. This limitation restricts the coverage and application of domain-specific knowledge, underscoring the need to develop larger, more versatile domain-specific LLMs for enhanced performance.

8 Potential Risks

The use of AI in synthetic chemistry carries several potential risks. One major concern is the possibility of misuse to produce dangerous or illicit substances, posing significant safety and ethical challenges. Additionally, inaccuracies in the generated content could lead to hazardous chemical reactions if not carefully verified, potentially causing harm or equipment damage. Over-reliance on AI-generated synthesis procedures without proper validation increases the risk of accidents and unsafe practices. Strict oversight and robust ethical guidelines are essential to mitigate these risks and ensure safe application

9 Acknowledgements

This project was supported by Shenzhen Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone, under Grant No. HTHZQSWS-KCCYB-2023052. .

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Derek T. Ahneman, Jesús G Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. 2018. Predicting reaction performance in C–N cross-coupling using machine learning. *Science*, 360(6385):186–190.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Alstonlo, Mario Krenn, Seyone Chithrananda, Andrew White, Florian Häse, Nathan Frey, Jannis Born, Andrei Voinea, Akshat Nigam, Darren Wee, François Bérenger, Haydn Jones, and Jocelyn. 2024. *aspuru-guzik-group/selfies*. GitHub.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A frontier large vision-language model with versatile abilities. *Preprint*, arXiv:2308.12966.

Guy W. Bemis and Mark A. Murcko. 1996. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39 15:2887–93.

Terra Blevins and Luke Zettlemoyer. 2022. Language contamination helps explains the cross-lingual capabilities of English pretrained models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574.

He Cao, Zijing Liu, Xingyu Lu, Yuan Yao, and Yu Li. 2023. InstructMol: Multi-modal integration for building a versatile and reliable molecular assistant in drug discovery. *Preprint*, arXiv:2311.16208.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian

Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality.

Davide Chicco, Valery V. Starovoitov, and Giuseppe Jurman. 2021. The benefits of the matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access*, 9:47112–47124.

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

Hanjun Dai, Chengtao Li, Connor W. Coley, Bo Dai, and Le Song. 2019. Retrosynthesis prediction with conditional graph logic network. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Rafael Delgado and Juan D. Núñez-González. 2019. Enhancing confusion entropy (CEN) for binary and multiclass classification. *PLoS ONE*, 14(1):e0210264.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Benchmark probing: Investigating data leakage in large language models. In *NeurIPS 2023 Workshop on Backdoors in Deep Learning - The Good, the Bad, and the Ugly*.

Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. 2002. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2Mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 595–607.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2024a. Mol-Instructions: A large-scale biomolecular instruction dataset for large language models. In *ICLR*. OpenReview.net.

Yin Fang, Ningyu Zhang, Zhuo Chen, Lingbing Guo, Xiaohui Fan, and Huajun Chen. 2024b. Domain-agnostic molecular generation with chemical feedback. In *The Twelfth International Conference on Learning Representations*.

H.A. Favre and W.H. Powell. 2014. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names*. International Union of Pure and Applied Chemistry. Royal Society of Chemistry.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Edward A. Hill. 1900. On a system of indexing chemical literature; adopted by the classification division of the U. S. patent office. *Journal of the American Chemical Society*, 22:478–494.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D. Bocarsly, Andres M. Bran, Stefan Bringuier, L. Catherine Brinson, Kamal Choudhary, Defne Circi, Sam Cox, Wibe A. de Jong, Matthew L. Evans, Nicolas Gastellu, Jerome Genzling, María Victoria Gil, Ankur K. Gupta, Zhi Hong, Alishba Imran, Sabine Kruschwitz, Anne Labarre, Jakub Lála, Tao Liu, Steven Ma, Sauradeep Majumdar, Garrett W. Merz, Nicolas Moitessier, Elias Moubarak, Beatriz Mouriño, Brenden Pelkie, Michael Pieler, Mayk Caldas Ramos, Bojana Ranković, Samuel G. Rodriques, Jacob N. Sanders, Philippe Schwaller, Marcus Schwarting,

Jiale Shi, Berend Smit, Ben E. Smith, Joren Van Herck, Christoph Völker, Logan Ward, Sean Warren, Benjamin Weiser, Sylvester Zhang, Xiaoqi Zhang, Ghezal Ahmad Zia, Aristana Scourtas, K. J. Schmidt, Ian Foster, Andrew D. White, and Ben Blaiszik. 2023. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2:1233–1250.

Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: Investigating the design space of visually-conditioned language models. *Preprint*, arXiv:2402.07865.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Y. Zaslavsky, Jian Zhang, and Evan E. Bolton. 2022. PubChem 2023 update. *Nucleic acids research*.

Mario Krenn, Florian Hase, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2019. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1.

Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, and Seokho Kang. 2022. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics*, 14:2.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello and-Sriniker, Gedeck, Gareth Jones, Nadine Schneider, Eisuke Kawashima, Dan Nealschneider, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Rachel Walker, Kazuya Ujihara, Daniel Probst, Guillaume Godin, Axel Pahl, Tadhurst-cdd, Juuso Lehtivarjo, Francois Berenger, and Jason D Biggs. 2024. RDKit: Open-source cheminformatics and machine learning.

V. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707.

Changmao Li and Jeffrey Flanigan. 2024. Task contamination: Language models may not be few-shot anymore. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18471–18480.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. Towards 3D molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations*.

Youwei Liang, Ruiyi Zhang, Li Zhang, and Pengtao Xie. 2023. DrugChat: Towards enabling ChatGPT-like capabilities on drug molecule graphs. *Preprint*, arXiv:2309.03907.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, volume 36.

Pengfei Liu, Yiming Ren, Jun Tao, and Zhixiang Ren. 2024c. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171:108073.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2023a. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training molecular graph representation with 3D geometry. In *International Conference on Learning Representations*.

Zequn Liu, Wei Zhang, Yingce Xia, Lijun Wu, Shufang Xie, Tao Qin, Ming Zhang, and Tie-Yan Liu. 2023b. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*.

Zhiyuan Liu, Sihang Li, Yancheng Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15623–15638.

Zhiyuan Liu, Yaorui Shi, An Zhang, Sihang Li, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024d. ReactXT: Understanding molecular "reaction-ship" via reaction-contextualized molecule-text pretraining. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023d. Rethinking tokenizer and decoder in masked graph modeling for molecules. In *NeurIPS*.

Zhiyuan Liu, An Zhang, Hao Fei, Enzhi Zhang, Xiang Wang, Kenji Kawaguchi, and Tat-Seng Chua. 2024e. ProtT3: Protein-to-text generation for text-based protein understanding. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Micha Livne, Zulfat Miftahutdinov, E. Tutubalina, Maksim Kuznetsov, Daniil Polykovskiy, Annika Brundyn, Aastha Jhunjhunwala, Anthony Costa, Alex Aliper, and Alex Zhavoronkov. 2023. nach0: multimodal natural and chemical languages foundation model. *Chemical Science*, 15:8380 – 8389.

Daniel Lowe. 2017. Chemical reactions from US patents (1976–Sep 2016).

Jieyu Lu and Yingkai Zhang. 2022. Unified deep learning model for multitask reaction predictions with explanation. *Journal of chemical information and modeling*.

Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. 2023a. One transformer can understand both 2D & 3D molecular data. In *The Eleventh International Conference on Learning Representations*.

Yanchen Luo, Junfeng Fang, Sihang Li, Zhiyuan Liu, Jiancan Wu, An Zhang, Wenjie Du, and Xiang Wang. 2024. Text-guided small molecule generation via diffusion model. *iScience*.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. 2023b. MolFM: A multimodal molecular foundation model. *Preprint*, arXiv:2307.09484.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023c. BioMedGPT: Open multimodal generative pre-trained transformer for biomedicine. *Preprint*, arXiv:2308.09442.

Kelong Mao, Xi Xiao, Tingyang Xu, Yu Rong, Junzhou Huang, and Peilin Zhao. 2021. Molecular graph enhanced transformer for retrosynthesis prediction. *Neurocomputing*, 457:193–202.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. 2024. MM1: Methods, analysis & insights from multimodal LLM pre-training. *Preprint*, arXiv:2403.09611.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. 2024. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. *Preprint*, arXiv:2402.17810.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. 2023. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Damith Perera, Joseph W. Tucker, Shalini Brahmbhatt, Christopher J. Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W. Sach. 2018. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. 2022. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital Discovery*, 1:91–97.

Yujie Qian, Zhening Li, Zhengkai Tu, and Connor W. Coley an Regina Barzilay. 2023. Predictive chemistry augmented with text retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12731–12745.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems*, volume 33.

Nadine Schneider, Roger A. Sayle, and Gregory A. Landrum. 2015. Get your atoms in order—an open-source implementation of a novel and robust molecular canonicalization algorithm. *Journal of Chemical Information and Modeling*, 55(10):2111–2120.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A. Hunter, Costas Bekas, and Alpha A. Lee. 2019. Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, page 1572–1583.

Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H. Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. 2020. Predicting retrosynthetic pathways using transformer-based models and a hypergraph exploration strategy. *Chemical Science*, page 3316–3325.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. 2021a. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152.

Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. 2021b. Prediction

of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2(1):015016.

Marwin HS Segler, Mike Preuss, and Mark P Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610.

Marwin HS Segler and Mark P Waller. 2017. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Vignesh Ram Somnath, Charlotte Bunne, Connor W. Coley, Andreas Krause, and Regina Barzilay. 2021. Learning graph models for retrosynthesis prediction. In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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

Mujeen Sung, Minbyul Jeong, Yonghwa Choi, Donghyeon Kim, Jinhyuk Lee, and Jaewoo Kang. 2022. BERN2: an advanced neural biomedical named entity recognition and normalization tool. *Bioinformatics*, 38(20):4837–4839.

Xiangru Tang, Howard Dai, Elizabeth Knight, Fang Wu, Yunyang Li, Tianxiao Li, and Mark Gerstein. 2024a. A survey of generative AI for de novo drug design: New frontiers in molecule and protein generation.

Xiangru Tang, Andrew Tran, Jeffrey Tan, and Mark Gerstein. 2024b. MolLM: A unified language model to integrate biomedical text with 2D and 3D molecular representations. *Bioinformatics*.

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

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Zhengkai Tu and Connor W. Coley. 2021. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Preprint*, arXiv:2110.09681.

Yue Wan, Chang-Yu Hsieh, Ben Liao, and Shengyu Zhang. 2022. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 22475–22490. PMLR.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, page 429–436, New York, NY, USA. Association for Computing Machinery.

Xiaorui Wang, Chang-Yu Hsieh, Xiaodan Yin, Jike Wang, Yuquan Li, Yafeng Deng, Dejun Jiang, Zhenxing Wu, Hongyan Du, Hongming Chen, Yun Li, Huanxiang Liu, Yuwei Wang, Pei Luo, Tingjun Hou, and Xiaojun Yao. 2023a. Generic interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research*, 6.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022b. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, pages 1–9.

Zeyuan Wang, Qiang Zhang, Keyan Ding, Ming Qin, Xiang Zhuang, Xiaotong Li, and Huajun Chen. 2023b. InstructProtein: Aligning human and protein language via knowledge instruction. *Preprint*, arXiv:2310.03269.

Zhengwei Wang, Yuxiao Wang, Xuan Zhang, Zhaoxu Meng, Zhenghe Yang, Wei Zhao, and Xuefeng Cui. 2022c. Graph-based reaction classification by contrasting between precursors and products. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 354–359.

David Weininger. 1988. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How powerful are graph neural networks? In *International Conference on Learning Representations*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? In *Thirty-Fifth Conference on Neural Information Processing Systems*.

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

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J. Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, María P. Magariños, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A. Patrícia Bento, Melissa F. Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. 2023. The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192.

Zheni Zeng, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2022. A deep-learning system bridging molecule structure and biomedical text with comprehension comparable to human professionals. *Nature communications*, 13(862).

Haiteng Zhao, Shengchao Liu, Chang Ma, Hannan Xu, Jie Fu, Zhi-Hong Deng, Lingpeng Kong, and Qi Liu. 2023. GIMLET: A unified graph-text model for instruction-based molecule zero-shot learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, Xin Chen, and Kai Yu. 2024. ChemDFM: Dialogue foundation model for chemistry. *Preprint*, arXiv:2401.14818.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-Mol: A universal 3D molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*.

A Data Collection

All the SMILES strings are canonicalized using RDKit (Landrum et al., 2024) to ensure a standard representation. We apply additional data cleaning steps, such as removing invalid SMILES and handling duplicate entries.

A.1 Data Cleaning

Data leakage in prior works. Our experiments identified data leakage issues in the previous popular benchmark study Mol-Instruction (Fang et al., 2024a). For example, in the retrosynthesis prediction task, we compared reactions in the train and test splits after canonicalizing SMILES and found that 72 chemical reactions in the test split also appeared in the train split. Moreover, in the reagent prediction task, 884 reactions in the train split were identical to those in the test split of the retrosynthesis prediction task. Additionally, the study employed a random split method for train and test sets, which resulted in significant molecular scaffold similarities (Fingerprint Tanimoto Similarity avg ~ 0.8) between the reactions in the train and test splits. Consequently, the test results on this benchmark lack generalizability for real-world applications.

Our non-overlapping, scaffold-based dataset splits. When splitting the dataset, we followed two principles: (1) Ensure that chemical reactions in the test splits of all downstream synthetic chemistry tasks do not appear in any train datasets, including both the pretraining and SFT train datasets; (2) Resample the test set based on a scaffold splitting approach, using a scaffold similarity threshold (Fingerprint Tanimoto Similarity set between 0.5 and 0.6). The number of samples was maintained consistent with the Mol-Instruction test set, with additional samples selected from the LlaSMol (Yu et al., 2024) test set. Figure 4 illustrates the scaffold similarity distribution of reaction SMILES between previous works and our resampled test set.

A.2 Data Collection and Preprocessing of PRESTO

In this section, we provide details on the data collection and preprocessing procedures for PRESTO two pretraining stages.

PubChem Caption Dataset for Mol-Text Alignment. We constructed a molecule caption dataset to enable the LLM to integrate molecule structure information and biomolecular domain knowledge during the initial alignment phase. Using the PubChem (Kim et al., 2022) database as the data source, we followed the construction procedures outlined in Liu et al. (2023a). For each molecule, we used the "description" field from its annotation page as the corresponding text description. This resulted in a dataset of 326,675 molecule-text pairs.

Interleaved Dataset for Domain Incremental Pretrain. Both BioT5 (Pei et al., 2023) and MolXPT (Liu et al., 2023b) use interleaved corpora, but they only replace identified molecule entities in the text with their corresponding 1D sequential representations. We take this a step further by replacing entities with encoded graph tokens. Motivationally, BioT5 masks parts of the 1D tokens to use a mask learning mechanism that promotes the model's learning of molecular sequential representation. MolXPT aims to enable molecular SMILES to leverage information from surrounding text and vice versa. Our approach takes a step further, it not only aims to align molecule tokens with text tokens but also to encourage interactions between multiple molecule entities. This is foundational for downstream tasks involving multi-molecule interactions, such as forward reaction prediction.

In detail, we compiled the interleaved molecule-text dataset primarily from USPTO-Applications (Lowe, 2017), consisting of approximately 2 million reactions and their corresponding application records published by USPTO between 2001 and September 2016. Raw XML files were downloaded, and key information for each reaction, including chemical reaction equations and textual descriptions of experimental procedures, was extracted. Following initial deduplication and filtering procedures outlined in (Wang et al., 2023a), we initially collected 1,593,329 procedure samples. Subsequently, we proceeded with two main preprocessing steps:

- Entity Recognition: We used the Named Entity Recognition tool BERN2 (Sung et al., 2022) to extract molecule entities from procedure paragraphs, retaining samples containing identifiable molecule entities. All extracted molecules' IUPAC names were then converted to SMILES format, suitable for further encoding into 2D molecular graphs. After this step, 1,592,462 samples remained.

- Removal of samples with excessive molecule entities and sequence length: To manage token

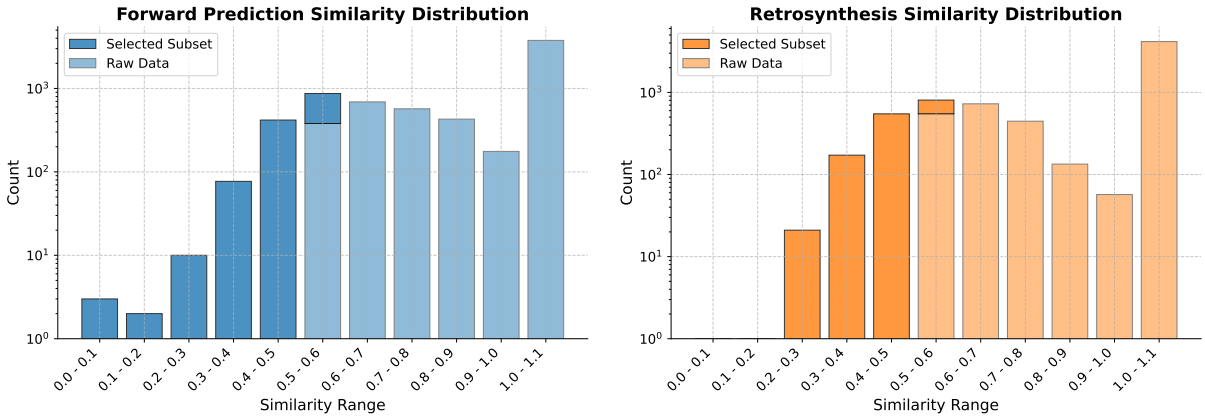


Figure 4: **Comparison of similarity distributions for reaction prediction datasets.** The plots show the count of scaffolds within each similarity range for the full test datasets provided in Yu et al. (2024) and Fang et al. (2024a) (raw data, lighter shade) and the selected subsets of 1000 scaffolds with the lowest similarities (darker shade).

space and prevent overly long sequences, samples containing more than 20 entities (filtering out 1,556 samples) and text sequences exceeding 1024 tokens (filtering out 2,197 samples) were removed. Finally, our constructed interleaved dataset comprises 1,588,709 samples, encompassing over 342,401 unique molecules. The statistics of the interleaved molecule-text dataset are shown in Figure 5.

Name Conversion Dataset for Domain Incremental Pretrain. We collected molecule entries from PubChem (Kim et al., 2022) and utilized the existing dataset from LLaSMol (Yu et al., 2024). LLaSMol originally presents four tasks: SMILES to Formula, SMILES to IUPAC name, IUPAC name to SMILES, and IUPAC name to Formula. We retained the latter two tasks as text-only data. To integrate molecule graph tokens into PRESTO, we replaced SMILES with graph representations using Landrum et al. (2024), creating two new tasks: Molecule Graph to Formula and Molecule Graph to IUPAC. Additionally, we derived a fifth task, Molecule Graph to SMILES, directly from the Kim et al. (2022) molecule entries by parsing the SMILES into graph representations similarly.

A.3 Downstream Tasks Dataset Construction

In this section, we provide details on the data collection process for all downstream tasks of PRESTO introduced in Section 3.4. Additionally, Table 5 provides a comprehensive comparison of the capabilities of each method across these tasks.

Reaction Prediction. We use USPTO-500-MT (Lu and Zhang, 2022; Fang et al., 2024a) and USPTO-full (Lowe, 2017; Yu et al., 2024) datasets for reaction prediction. The training set of Fang et al. (2024a) has been chosen for its wide usage (Pei et al., 2023, 2024; Livne et al., 2023; Cao et al., 2023; Zhao et al., 2024). However, while several previous works have reported near-optimal accuracy on the test set of Fang et al. (2024a), we argue that most models still fail in real-world hard cases. To enhance the original test set's complexity, we add more challenging cases from Yu et al. (2024)'s test set based on Bemis-Murcko scaffolds (Bemis and Murcko, 1996). This ensures lower similarity between train and test sets. The new test set has 1,000 samples to thoroughly evaluate the model's generalization ability.

Reaction Condition Prediction. The reaction condition prediction tasks use combined data from TextReact (Qian et al., 2023) and Mol-Instruction (Fang et al., 2024a), both sourced from the USPTO dataset. Following Qian et al. (2023), we further annotate reaction condition prediction into subtasks with reagents, catalysts, and solvents. Notably, 65.75% of the training reactions and 68.47% of the test reactions in Qian et al. (2023) overlap with Fang et al. (2024a). To ensure fair comparison and utilize the additional data, we create a new dataset by combining the overlapping reactions. The data is split into train/valid/test sets with a ratio of 8:1:1 for each task.

Reagent Selection. Our study utilizes the reagent selection dataset from ChemLLMBench (Guo et al., 2023), comprising 4,255 valid samples originally sourced from the Suzuki High-Throughput Experimentation (HTE) dataset (Perera et al., 2018). Each

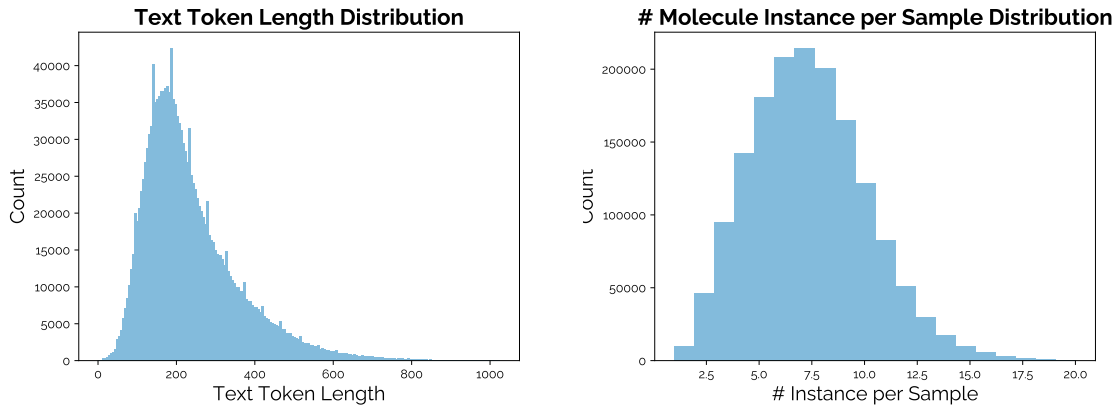


Figure 5: **Statistics of the Interleaved Molecule-Text Dataset.**

Method	Forward	Retro	Reaction Condition Pred				Reagent Recommend	Reaction Type	Yield
			All	Reagent	Catalyst	Solvent			
T5Chem (Lu and Zhang, 2022)	✓	✓	✓	✗	✗	✗	✗	✓	✓
Text+ChemT5 (Christofidellis et al., 2023)	✓	✓	✗	✗	✗	✗	✗	✗	✗
TextReact (Qian et al., 2023)	✗	✓	✓	✓	✓	✓	✓	✗	✗
ChemDFM (Zhao et al., 2024)	✓	✓	✓	✗	✗	✓	✓	✗	✓
Mol-Instruction (Fang et al., 2024a)	✓	✓	✓	✗	✗	✗	✗	✗	✗
LlaSMol (Yu et al., 2024)	✓	✓	✗	✗	✗	✗	✗	✗	✗
BioT5+ (Pei et al., 2024)	✓	✓	✓	✗	✗	✗	✗	✗	✗
InstructMol (Cao et al., 2023)	✓	✓	✓	✗	✗	✗	✗	✗	✗
nach0 (Livne et al., 2023)	✓	✓	✓	✗	✗	✗	✗	✗	✗
PRESTO	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 5: **Comparison of various models across different chemical reaction prediction tasks.** The table summarizes the capabilities of each method in forward reaction prediction, retrosynthesis prediction, reaction condition prediction (overall, reagent, catalyst, and solvent), reagent recommendation, reaction type prediction, and yield prediction. PRESTO demonstrates comprehensive support across all tasks.

sample includes reactants, a product, and a list of candidate reagents. The objective is to select the most suitable reagent from the candidate list to facilitate the reaction. The dataset is divided into 3,955 training samples and 300 testing samples, maintaining the same test split as Guo et al. (2023).

Reaction Type Classification. For reaction type classification, we use the USPTO 1K TPL dataset (Schwaller et al., 2021a) derived from the USPTO patent database (Lowe, 2017), which contains 445,115 reactions labeled with 1000 reaction classes. Keeping the original configuration, the dataset is split into 360,545 samples for training, 40,059 for validation, and 44,511 for testing.

Yield Regression. In this task, we use the Buchwald-Hartwig dataset (Ahneman et al., 2018) and the Suzuki-Miyaura dataset (Perera et al., 2018) collected from Schwaller et al. (2021b). The Buchwald-Hartwig dataset contains 3,955 reactions, while the Suzuki-Miyaura dataset contains 5,760 reactions. We follow the approach of ChemLLMBench, using their predefined test sets (100 tests each). Notably, we convert it into a regression task, and the yield values are normalized to the range [0, 1].

A.4 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 PRESTO. 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

B.1 Evaluation Metrics

We utilize a variety of metrics to comprehensively evaluate the performance of the models across different types of tasks. The key metrics used for each

DATA SOURCES	LICENSE URL	LICENSE NOTE
PubChem	https://www.nlm.nih.gov/web_policies.html	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	https://creativecommons.org/licenses/by/4.0/	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.
IUPAC	https://iupac.org/wp-content/uploads/2021/06/iupac-inchi-license_2020.pdf	An "IUPAC license" generally refers to the permissions, guidelines, or rights associated with using the standards, software, data, or publications provided by the International Union of Pure and Applied Chemistry (IUPAC). This can include adhering to IUPAC's chemical nomenclature guidelines in scientific communication, using their proprietary software or databases under specific licensing terms, and obtaining permissions to reproduce or adapt copyrighted materials.
USPTO	https://www.uspto.gov/learning-and-resources/open-data-and-mobility	It can be freely used, reused, and redistributed by anyone.

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

type of task are as follows.

Classification Tasks. For classification tasks, we report the following metrics:

- **Accuracy**: The ratio of correctly classified samples.
- **CEN** (Delgado and Núñez-González, 2019): The CEN score is a measure of the overall entropy of a confusion matrix, which is used to evaluate classifiers in multi-class problems.
- **MCC** (Chicco et al., 2021): The MCC score is a balanced measure of binary classification quality, considering true and false positives and negatives.

Regression Tasks. For regression tasks, we consider the following metrics:

- **MAE**: Mean Absolute Error, the average absolute difference between predicted and actual values.
- **MSE**: Mean Squared Error, the average squared difference between predicted and actual values.
- R^2: The coefficient of determination, indicating the proportion of variance in the target variable that is predictable from the input features.

Molecule Generation Tasks. For tasks involving SMILES (Weininger, 1988) representations of molecules, we calculate:

- **Exact Match**: The proportion of predicted SMILES strings that exactly match the ground truth after canonicalization.
- **BLEU** (Papineni et al., 2001): The BLEU score treats the SMILES strings as text, measuring n-gram overlap between predictions and references.
- **Levenshtein Distance** (Levenshtein, 1966): The minimum number of single-character edits required to change the predicted SMILES into the reference.
- **RDKit Similarity** (Landrum et al., 2024): The Tanimoto similarity between RDKit fingerprints of the predicted and reference molecules.
- **MACCS Keys Similarity** (Durant et al., 2002): The Tanimoto similarity between MACCS keys fingerprints of the molecules.
- **Morgan Fingerprint Similarity** (Schneider et al., 2015): The Tanimoto similarity between Morgan circular fingerprints of the molecules.
- **Validity**: The proportion of predicted SMILES strings that can be successfully parsed into valid molecule structures by RDKit.

Note that if the origin model is trained on SELFIES (Krenn et al., 2019), we use Alstonlo et al. (2024) to translate the generated SELFIES to SMILES before evaluation.

B.2 Experimental Details

Here we detail the hyperparameters for PRESTO pretraining and SFT.

PRESTO Alignment Stage. We employed the PubChem molecule caption dataset, comprising approximately 327K samples, for training over 5 epochs. Training was conducted using 8×A6000 GPUs, with a total batch size of 128. AdamW optimizer was utilized with $\beta = (0.9, 0.999)$ and a learning rate of 2e-3, without weight decay. The learning rate was initially warmed up over 3% of the total training steps, followed by a cosine decay schedule. The model's maximum sequence length was set to 2048 for the base LLM. To conserve

CUDA memory, we employed DeepSpeed ZeRO-2 strategy and gradient checkpointing.

PRESTO Domain Incremental Pretrain Stage. Using the projector checkpoint from the alignment stage, training followed the fundamental settings of the alignment stage, with adjustments made to the total batch size, set to 64, and the learning rate, set to 2e-5. Due to the prohibitive costs associated with fully finetuning the base 7B LLMs and the extensive pretraining dataset, all experiments were limited to one epoch.

Supervised Finetuning. We utilize the updated projector and LLM weights from the pretraining stage and combine all downstream task training sets for joint model training. For the full finetuning experiment, we train for three epochs by default, using the same hyperparameters as in the pretraining stage except for setting the total batch size to 128. For the LoRA ablation, we set the peak learning rate to 8e-5.

C More Ablations

This section extends Section 4 to introduce more findings according to the ablation experiments.

C.1 Analyzing SFT

Here, we explore important aspects of supervised finetuning, such as parameters, training time, and data scaling.

Finding 5: Updating LLMs is essential. We conducted an ablation study on the trainable parameters of LLMs during the SFT stage (Figure 6a), progressing from not updating any LLM parameters to updating the attention block's `q_proj` and `v_proj` layers with LoRA, then updating all linear layers except the `lm_head` layer with LoRA, and finally fully finetuning all parameters. All experiments involved training for 3 epochs on the SFT dataset. We found that not updating the LLM parameters during SFT led to nearly zero performance, highlighting the necessity of parameter updates for adapting to downstream tasks. Incorporating LoRA modules significantly boosted performance, and adding more trainable LoRA modules consistently improved results. Moreover, when computational resources allow, full-tuning outperforms LoRA-tuning across various downstream tasks.

Finding 6: Balancing SFT training time optimizes downstream task performance. We investigate the impact of SFT training time on a subset of our SFT training dataset (1/7 size, detailed in the Appendix). Unlike existing Vision LMs, which typically undergo only one epoch of training, we compare performance across different numbers of epochs. We observe severe underfitting with only one epoch of training. Surprisingly, we find steady improvement across all tasks when trained for up to three epochs but encounter overfitting when training to four epochs, leading to performance degradation. In conclusion, we recommend training for three epochs for optimal performance on downstream tasks.

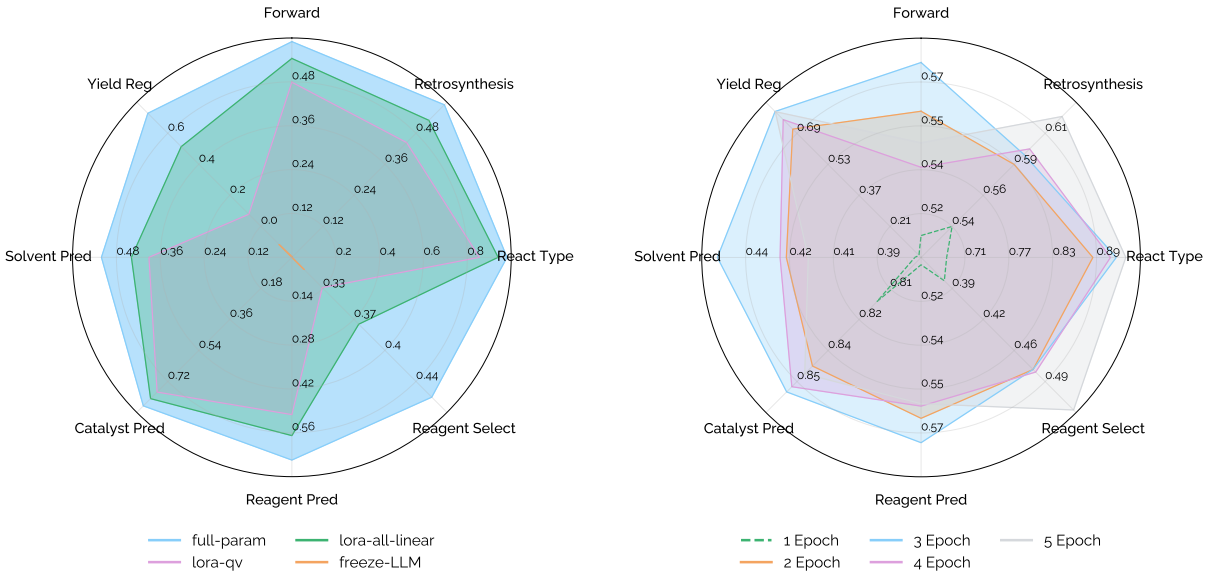
Finding 7: Coverage and diversity of SFT dataset are critical for better results. We examined the impact of data repetition (i.e., allocating FLOPs across multiple epochs on the same data) and SFT-data size on downstream tasks. In our experiments on forward and retrosynthesis prediction, we fixed the training FLOPs (equivalent to the FLOPs used to train for 1 epoch with the full dataset) and successively halved the training dataset while doubling the number of training epochs. We used two subsampling methods: (1) random subsampling and (2) hierarchical subsampling based on scaffold clustering. Figure 7 revealed that for a fixed compute budget, training up to four epochs with repeated data resulted in negligible changes in loss compared to using unique data. Moreover, we found that the coverage and diversity of the SFT training set are crucial; even when the training set size was halved, maintaining the number of scaffold clusters led to higher performance on the test set.

C.2 Graph v.s. SMILES

Finding 8: 2D graphs outperform 1D SMILES in modeling molecules. We provide a comparison between Vicuna v1.5-7B (using 1D SMILES as molecule input) and PRESTO (using 2D graphs as molecule input) across several synthetic chemistry tasks. To ensure fairness, we bypassed PRESTO's two pretraining stages. Table 7 and 8 show that the 2D graph modality outperforms the 1D sequential representation for modeling molecules across synthetic chemistry tasks.

C.3 Generalize to Smaller LLMs

A straightforward question arises: does adapting the PRESTO method to smaller LLMs still yield improvements? To investigate this, we implement PRESTO training on phi-3-mini (Abdin et al.,



(a) **# Trainable Param Ablation**

(b) **Scaling SFT Train Time**

Figure 6: **Performance analysis of different training strategies and dataset configurations.** (a) Ablation study on the trainable parameters in the LLM during SFT. An increase in trainable parameters consistently enhances performance. (b) Analysis of training duration impacts on SFT. Performance improves up to three epochs, while training for four epochs results in overfitting.

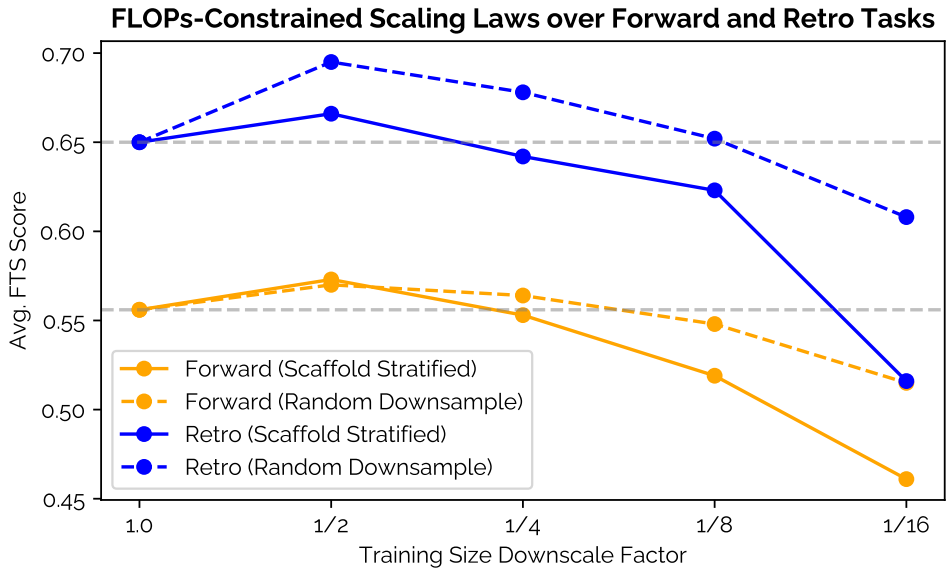


Figure 7: **Impact of SFT training dataset coverage and diversity on downstream task performance.** Training up to four epochs with repeated data resulted in negligible changes in loss compared to using unique data. Maintaining the number of scaffold clusters even when the training set size was halved led to higher performance on the test set.

2024) (3.8B) and assess its efficacy in forward reaction prediction, retrosynthesis, and condition prediction. Our findings indicate that PRESTO enhances phi-3-mini's performance across all tasks, surpassing the best baseline for each. For illustration, we present phi-3-mini's performance on forward reaction prediction as in Table 9.

D Instruction Templates

In this section, we provide a basic description of the instruction templates utilized in PRESTO. These templates are designed to guide the model during pretraining and downstream tasks. We have a variety of templates for each task, and we present a randomly selected template in this part.

MODEL	EXACT↑	BLEU↑	LEVENSHTEIN↓	RDK FTS↑	MACCS FTS↑	MORGAN FTS↑	VALIDITY↑
Forward Reaction Prediction							
Vicuna v1.5-7b	0.250	0.659	12.506	0.513	0.600	0.495	0.903
ours (w/o stage-1&2)	0.298	0.763	12.763	0.576	0.663	0.557	0.975
Retrosynthesis Prediction							
Vicuna v1.5-7b	0.220	0.756	15.692	0.658	0.758	0.616	0.943
ours (w/o stage-1&2)	0.184	0.896	16.393	0.681	0.766	0.627	0.959
Reaction Condition Prediction (Reagent)							
Vicuna v1.5-7b	0.315	0.585	8.664	0.576	0.587	0.473	1.0
ours (w/o stage-1&2)	0.405	0.747	6.940	0.642	0.651	0.556	1.0
Reaction Condition Prediction (Catalyst)							
Vicuna v1.5-7b	0.685	0.703	2.451	0.883	0.869	0.692	1.0
ours (w/o stage-1&2)	0.748	0.822	1.851	0.917	0.899	0.752	1.0
Reaction Condition Prediction (Solvent)							
Vicuna v1.5-7b	0.320	0.436	3.809	0.459	0.486	0.427	1.0
ours (w/o stage-1&2)	0.366	0.662	2.948	0.487	0.507	0.461	0.912

Table 7: Comparison of 1D SMILES and 2D graphs as representations of molecules for tasks involving forward reaction prediction, retrosynthesis prediction, and reaction condition prediction.

METHOD	REACTANT	SOLVENT	LIGAND
Reagent Selection			
Vicuna v1.5-7b	0.78	0.58	0.44
ours (w/o stage-1&2)	0.78	0.55	0.63

METHOD	ACC↑	CEN↓	MCC↑
Reaction Type Classification			
Vicuna v1.5-7b	0.888	0.048	0.887
ours (w/o stage-1&2)	0.988	0.007	0.987

METHOD	B-H	S-M
Yield Regression		
Vicuna v1.5-7b	-0.131	0.151
ours (w/o stage-1&2)	0.695	0.480

Table 8: Comparison of 1D SMILES and 2D graphs as representations of molecules for tasks involving reagent selection, reaction type classification, and yield regression.

Base LLMs	**w/ PRESTO Training**	**FTS↑**
nach0-base	No	0.642
Vicuna-v1.5-7B	No	0.536
Vicuna-v1.5-7B	Yes	**0.590** (+10.1%)
Phi-3-mini	No	0.560
Phi-3-mini	Yes	**0.644** (+15.0%)

Table 9: Fingerprint similarity for different scale of LLMs with and without PRESTO training on forward reaction prediction task.

D.1 Template for Pretraining

Here are six templates used in the pretraining stage of PRESTO:

1. PubChem Caption (Table 10)
2. IUPAC to Formula (Table 11)
3. IUPAC to SMILES (Table 12)
4. Molecule Graph to Formula (Table 13)
5. Molecule Graph to IUPAC (Table 14)
6. Molecule Graph to SMILES (Table 15)

D.2 Template for Downstream Tasks

Here are 10 templates used for downstream tasks of PRESTO:

1. Forward Prediction (Table 16)
2. Retrosynthesis Prediction (Table 17)
3. Catalyst Prediction (Table 18)
4. Reagent Prediction (Table 19)
5. Solvent Prediction (Table 20)
6. Reagent Selection (Table 21)
7. Ligand Selection (Table 22)
8. Solvent Selection (Table 23)
9. Yield Prediction (Table 24)
10. Reaction Type Classification (Table 25)

E Case Studies

We show some selected cases for forward prediction (Table 8), retrosynthesis prediction (Table 9), reagent prediction (Table 10), solvent prediction (Table 12), and catalyst prediction tasks (Table 11).

System Prompt:
You are a chemist. Now you are given a representation of a molecule. Please help me to understand the molecule.
Instruction:
Provide a brief overview of this molecule: <molecule>.
Output:
Sure! Here is a description of this molecule. <caption>.

Table 10: **An instruction template for generating captions for molecules.** This template guides the generation of a brief description or overview of the given molecule representation.

System Prompt:
You are a chemist. Please follow the instructions to convert the structure to the corresponding name.
Instruction:
<input> is the IUPAC name of a molecule. Please give its molecular formula.
Output:
The molecular formula is <output>.

Table 11: **An instruction template for converting IUPAC names to molecular formulas.** This template guides the conversion from the given IUPAC name to its corresponding molecular formula.

System Prompt:
You are a chemist. Please follow the instructions to convert the structure to the corresponding name.
Instruction:
Convert the IUPAC name of a molecule <input> into SMILES representation.
Output:
The SMILES representation is <output>.

Table 12: **An instruction template for converting IUPAC names to SMILES representations.** This template guides the conversion from the given IUPAC name to its corresponding SMILES representation.

System Prompt:
You are a chemist. Please follow the instructions to convert the structure to the corresponding name.
Instruction:
<input> is the representation of a molecule. What is its molecular formula?
Output:
The molecular formula is <output>.

Table 13: **An instruction template for converting molecular graph to molecular formula.** This template guides the conversion from the given graph representation to its corresponding molecular formula.

System Prompt:
You are a chemist. Please follow the instructions to convert the structure to the corresponding name.
Instruction:
<input> is the representation of a molecule. What is its IUPAC name?
Output:
The IUPAC name is <output>.

Table 14: **An instruction template for converting molecule graph to IUPAC name.** This template guides the conversion from the given graph representation to its corresponding IUPAC name.

System Prompt:
You are a chemist. Please follow the instructions to convert the structure to the corresponding name.
Instruction:
The representation of a certain molecule is <input>. Can you provide its SMILES representation?
Output:
The SMILES representation is <output>.

Table 15: **An instruction template for converting the molecule graph to SMILES representation.** This template guides the conversion from the given graph representation to its corresponding SMILES representation.

System:
You are a chemist. Your task is to predict the SMILES representation of the product molecule, given the molecule representations of the reactants.
Instruction:
Using <reactant_1>.<reactant_2>.<reactant_3> as the reactants and reagents, tell me the potential product.
Output:
Sure. A potential product: <product_1>.<product_2>.

Table 16: **An instruction template for forward prediction.** This template guides the prediction of the product based on the given reactants and reagents. The reactants and reagents are specified, and the model must predict the potential product from the reaction.

System:
You are a chemist. Your task is to predict the SMILES representation of the reactant molecules, given the molecule representations of the product.
Instruction:
Using <product_1>.<product_2>.<product_3> as the products, predict the possible reactants that could have been utilized to synthesize these products.
Output:
Here are possible reactants: <reactant_1>.<reactant_2>.

Table 17: **An instruction template for retrosynthesis prediction.** This template guides the prediction of the possible reactants based on the given product. The product is specified, and the model must predict the reactants that could have been used to synthesize this product.

System Prompt:
You are a chemist. Now, you are given a reaction equation. Your task is to predict the SMILES representation of the catalyst, given molecule representation of the reaction.
Instruction:
Based on the given chemical reaction: <reactant_1>.<reactant_2>.<reactant_3> » <product_1>.<product_2>, propose some likely catalysts that might have been utilized.
Output:
A possible catalyst can be <catalyst>.

Table 18: **An instruction template for catalyst prediction.** This template guides the prediction of possible catalysts based on the given reaction components. The reactants and products are specified, and the model must predict the potential catalyst from the reaction.

System Prompt:
You are a chemist. Now, you are given a reaction equation. Your task is to predict the SMILES representation of the reagents, given molecule representation of the reaction.
Instruction:
Based on the given chemical reaction: <reactant_1>.<reactant_2>.<reactant_3> » <product_1>.<product_2>, propose some likely reagents that might have been utilized.
Output:
A possible reagent can be <reagent>.

Table 19: **An instruction template for reagent prediction.** This template guides the prediction of possible reagents based on the given reaction components. The reactants and products are specified, and the model must predict the potential reagent from the reaction.

System Prompt:
You are a chemist. Now, you are given a reaction equation. Your task is to predict the SMILES representation of the solvents, given molecule representation of the reaction.
Instruction:
Based on the given chemical reaction: <reactant_1>.<reactant_2>.<reactant_3> » <product_1>.<product_2>, propose some likely solvents that might have been utilized.
Output:
A possible solvent can be <solvent>.

Table 20: **An instruction template for solvent prediction.** This template guides the prediction of possible solvents based on the given reaction components. The reactants and products are specified, and the model must predict the potential solvent from the reaction.

System Prompt:
You are an expert chemist. Given one reactant, two reagents, and one solvent of a Suzuki reaction, predict the optimal reactant that maximizes the yield with the rest of the reaction components. Only return the option from the given list.
Instruction:
Given the rest of the reaction components: <reactant_1> > <reagent_1>.<reagent_2> » <solvent>. Select the optimal reactant: <reactant_2>.<reactant_3>
Output:
Optimal reactant: <reactant_3>.

Table 21: **An instruction template for reagent selection.** This template guides the prediction of the optimal reactant based on the given reaction components. The reactant, reagents, and solvent are specified, and the model must choose the best reactant from the provided list.

System Prompt:
You are an expert chemist. Given two reactants, one reagent, and one solvent of a Suzuki reaction, predict the optimal ligand that maximizes the yield with the rest of the reaction components. Only return the option from the given list.
Instruction:
Given the rest of the reaction components: <reactant_1>.<reactant_2> » <reagent>.<solvent>. Select the optimal ligand: <ligand_1>.<ligand_2>
Output:
Optimal ligand: <ligand_1>.

Table 22: **An instruction template for ligand selection.** This template guides the prediction of the optimal ligand based on the given reaction components. The reactants, reagents, and solvents are specified, and the model must choose the best ligand from the provided list.

System Prompt:
You are an expert chemist. Given two reactants, one ligand, and one base of a Suzuki reaction, predict the optimal solvent that maximizes the yield with the rest of the reaction components. Only return the option from the given list.
Instruction:
Given the rest of the reaction components: <reactant_1>.<reactant_2> » <ligand>.<base>. Select the optimal solvent: <solvent_1>.<solvent_2>
Output:
Optimal solvent: <solvent_2>.

Table 23: **An instruction template for solvent selection.** This template guides the prediction of the optimal solvent based on the given reaction components. The reactants, ligand, and base are specified, and the model must choose the best solvent from the provided list.

System Prompt:
You are a chemist. Now, you are given a reaction equation. Your task is to predict the yield ratio of the reaction. The return value should be in the range of 0-1. The higher the value, the more likely the reaction is to occur.
Instruction:
Based on the given chemical reaction: <reactant_1>.<reactant_2>.<reactant_3> » <product_1>.<product_2>, what is the yield ratio of the reaction?
Output:
The yield ratio is <ratio>.

Table 24: **An instruction template for yield prediction.** This template guides the prediction of the yield ratio based on the given reaction components. The reactants and products are specified, and the model must predict the yield ratio from the reaction.

System Prompt:
You are a chemist. Now, you are given a reaction equation. Your task is to predict the class of the reaction. Your task is to predict the class number of the reaction.
Instruction:
Based on the given chemical reaction: <reactant_1>.<reactant_2>.<reactant_3> » <product_1>.<product_2>, predict the class number of the reaction.
Output:
The class number is <class_number>.

Table 25: **An instruction template for reaction type classification.** This template guides the prediction of the reaction class number based on the given reaction components. The reactants and products are specified, and the model must predict the reaction class number from the reaction.

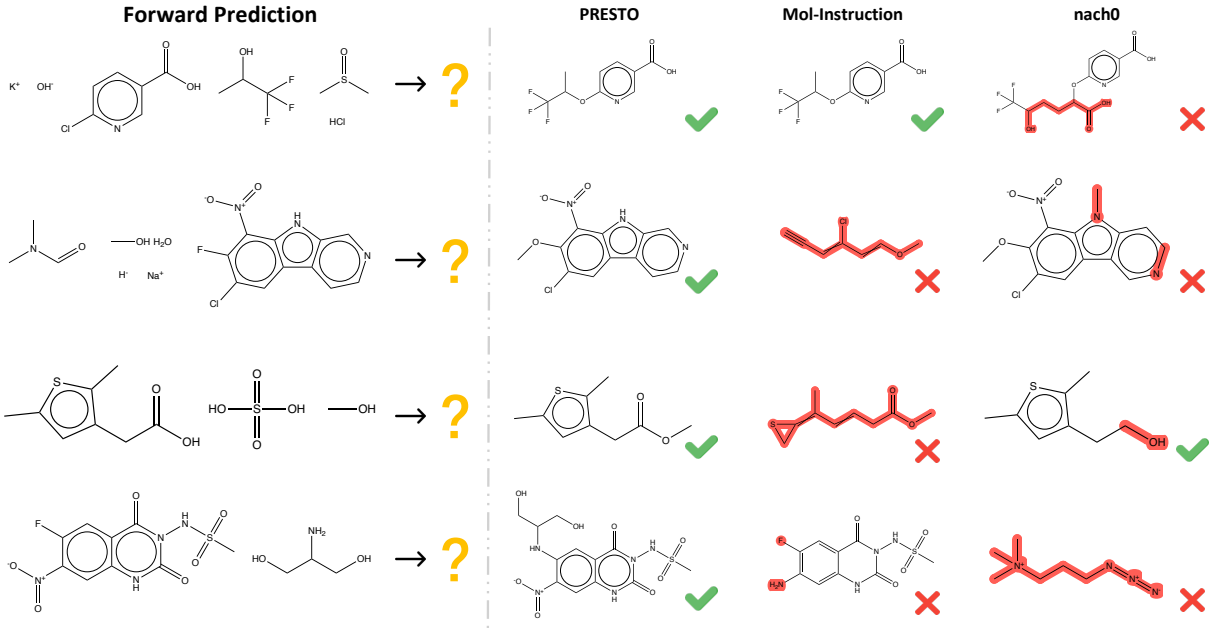


Figure 8: More examples of the **Forward Prediction** task. We include Mol-Instruction (Fang et al., 2024a) and nach0 (Livne et al., 2023) as baselines.

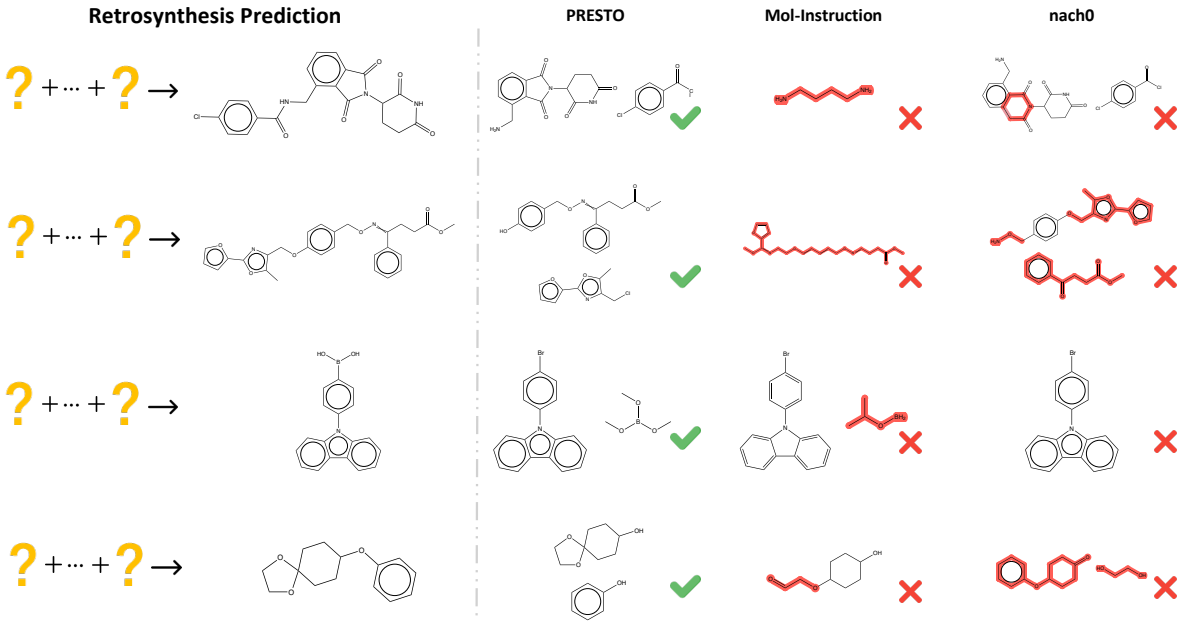


Figure 9: More examples of the **Retrosynthesis Prediction** task. We include Mol-Instruction (Fang et al., 2024a) and nach0 (Livne et al., 2023) as baselines.

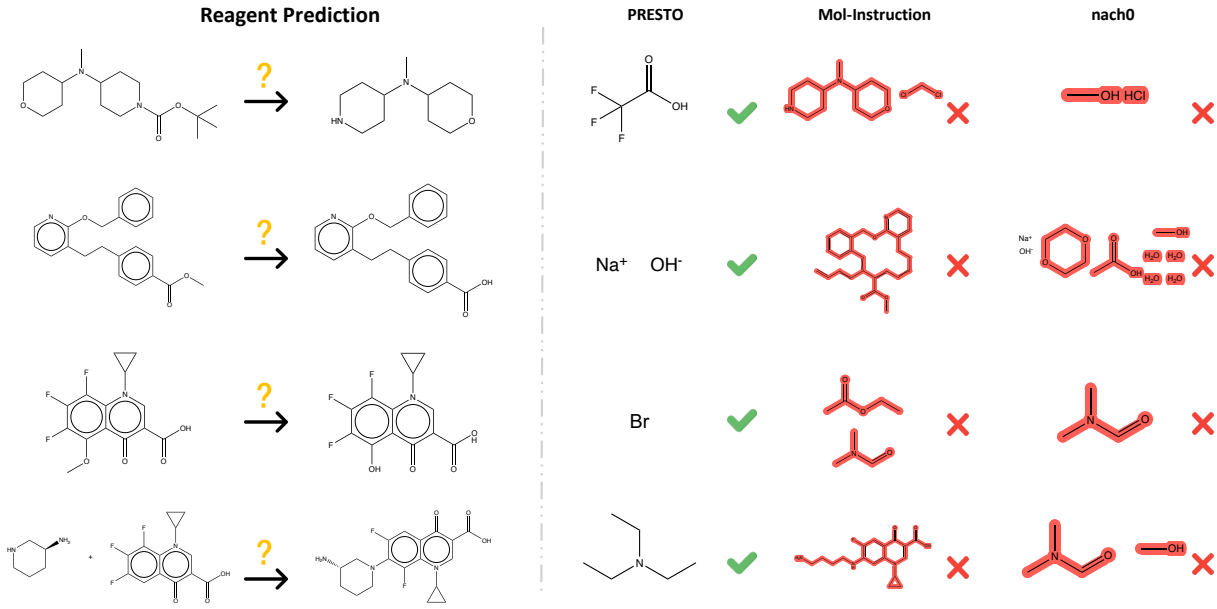


Figure 10: More examples of the **Reagent Prediction** task. We include Mol-Instruction (Fang et al., 2024a) and nach0 (Livne et al., 2023) as baselines.

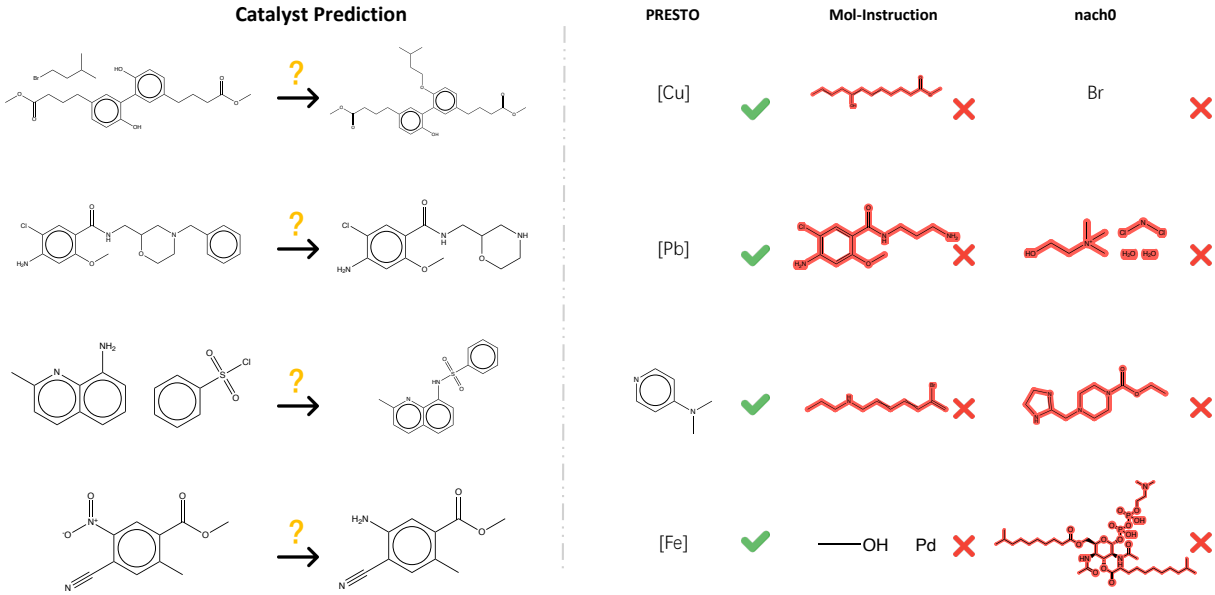


Figure 11: More examples of the **Catalyst Prediction** task. We include Mol-Instruction (Fang et al., 2024a) and nach0 (Livne et al., 2023) as baselines.

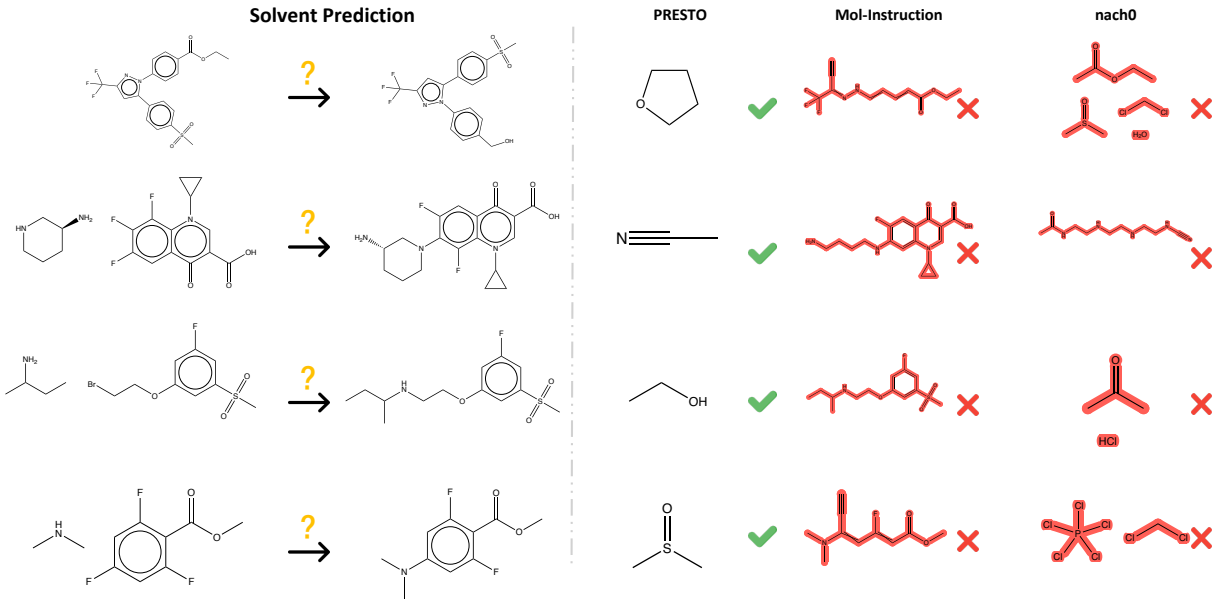


Figure 12: More examples of the **Solvent Prediction** task. We include Mol-Instruction (Fang et al., 2024a) and nach0 (Livne et al., 2023) as baselines.