ReactXT: Understanding Molecular "Reaction-ship" via Reaction-Contextualized Molecule-Text Pretraining

Zhiyuan Liu[1*] **Yaorui Shi**[2*] **An Zhang**[1] **Sihang Li**[2]
Enzhi Zhang[3] **Xiang Wang**[2†] **Kenji Kawaguchi**[1] **Tat-Seng Chua**[1]
[1]National University of Singapore
[2]University of Science and Technology of China [3]Hokkaido University
{acharkq,yaoruishi,an.zhang3.14,sihang0520,xiangwang1223}@gmail.com
enzhi.zhang.n6@elms.hokudai.ac.jp, {kenji,chuats}@comp.nus.edu.sg

Abstract

Molecule-text modeling, which aims to facilitate molecule-relevant tasks with a textual interface and textual knowledge, is an emerging research direction. Beyond single molecules, studying reaction-text modeling holds promise for helping the synthesis of new materials and drugs. However, previous works mostly neglect reaction-text modeling: they primarily focus on modeling individual molecule-text pairs or learning chemical reactions without texts in context. Additionally, one key task of reaction-text modeling – experimental procedure prediction – is less explored due to the absence of an open-source dataset. The task is to predict step-by-step actions of conducting chemical experiments and is crucial to automating chemical synthesis. To resolve the challenges above, we propose a new pretraining method, **ReactXT**, for reaction-text modeling, and a new dataset, **OpenExp**, for experimental procedure prediction. Specifically, ReactXT features three types of input contexts to incrementally pretrain LMs. Each of the three input contexts corresponds to a pretraining task to improve the text-based understanding of either reactions or single molecules. ReactXT demonstrates consistent improvements in experimental procedure prediction and molecule captioning and offers competitive results in retrosynthesis. Our code is available at https://github.com/syr-cn/ReactXT.

1 Introduction

Multi-modal large language models (LMs) have recently attracted extensive research attention. Remarkably, in the vision-language domain, LMs enhanced with visual encoders show impressive results in visual question-answering and image captioning (Liu et al., 2023a; Li et al., 2023). Inspired by their successes, molecule-text modeling (MTM) becomes an emerging research field (Liu et al., 2023b; Zeng et al., 2022; Su et al., 2022), aiming to build the natural language interface for molecular tasks, including text-guided molecule generation, molecule captioning, and molecule-text retrieval (Edwards et al., 2022; Liu et al., 2022).

Building upon these MTM works, we study reaction-text modeling (RTM), aiming to improve LMs' performance on reaction-relevant tasks. Chemical reactions, involving the transformation of reactants into products, are fundamental to advancing drug discovery and material science (Schwaller et al., 2022). Revisiting prior works, we identify key research gaps in both the learning paradigm and the evaluation benchmark for RTM:

- **Learning Paradigm.** Most prior works either focus on generating the textual description of a single molecule (*cf.* Figure 1a) (Liu et al., 2023b; Edwards et al., 2022; Su et al., 2022), or apply LMs for chemical reaction prediction without including the textual descriptions of molecules/reactions in the context (*cf.* Figure 1b) (Christofidellis et al., 2023; Fang et al., 2023; Born and Manica, 2023). Such methods overlook the potential knowledge in textual descriptions to improve performance. Pioneer works (Shi et al., 2023; Guo et al., 2023) include labels of molecular roles and experimental conditions when prompting ChatGPT, but achieve suboptimal performances for being limited to prompt engineering.

- **Evaluation Benchmark.** An open-source dataset for experimental procedure prediction is notably missing. As illustrated in Figure 2, experimental procedure prediction aims to deduce the

[*] Equal contribution.

[†] Corresponding author. Xiang Wang is also affiliated with Institute of Dataspace, Hefei Comprehensive National Science Center.

Findings of the Association for Computational Linguistics: ACL 2024, pages 5353–5377
August 11-16, 2024 ©2024 Association for Computational Linguistics

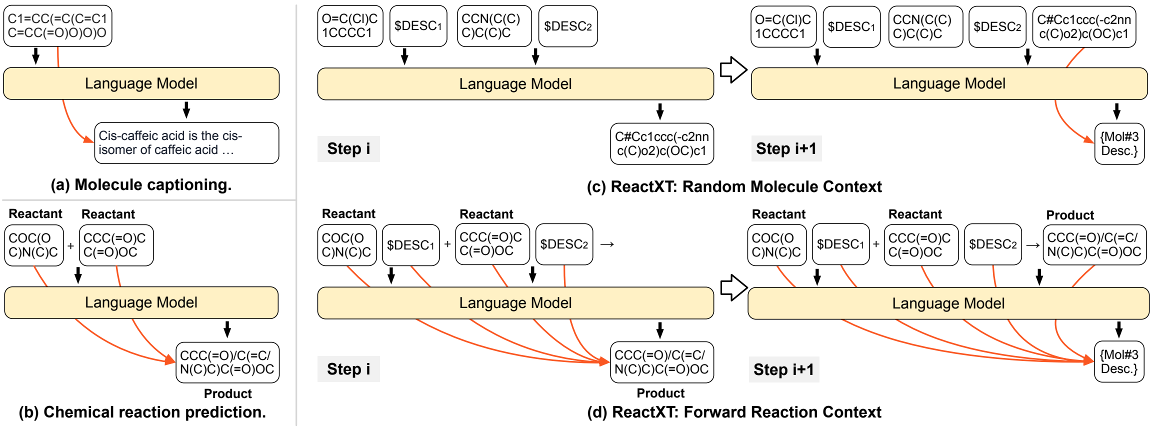
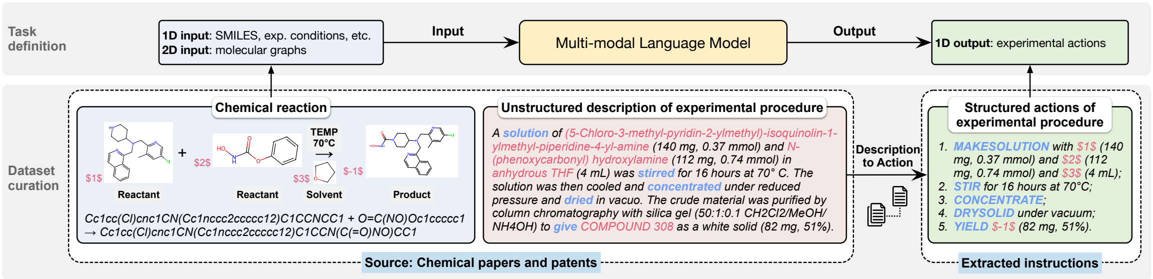


Figure 1: Comparison of molecule-text generative modeling methods. Orange arrows → denote the chemical relations for generation. 2D graph embeddings (Liu et al., 2023b) are omitted here for simplicity, but are added in the final framework for improved performance. $\$\mathtt{DESC}_j$ denotes the description of the j-th molecule. The chemical reaction in Figures (b) and (d) is: COC(OC)N(C)C + CCC(=O)CC(=O)OC → CCC(=O)/C(=C/N(C)C)C(=O)OC.


Task
definition
1D input: SMILES, exp. conditions, etc.
2D input: molecular graphs
Input
Multi-modal Language Model
Output
1D output: experimental actions
Dataset
curation
Chemical reaction
1
+
2
TEMP
70°C
3
-1
Reactant
Reactant
Solvent
Product
Unstructured description of experimental procedure
Description
to Action
Structured actions of
experimental procedure
Source: Chemical papers and patents
Extracted instructions

Cc1cc(Cl)cnc1CN(Cc1nccc2ccccc12)C1CCNCC1 + O=C(NO)Oc1ccccc1
→ Cc1cc(Cl)cnc1CN(Cc1nccc2ccccc12)C1CCN(C(=O)NO)CC1

A solution of (5-Chloro-3-methyl-pyridin-2-ylmethyl)-isoquinolin-1-ylmethyl-piperidine-4-yl-amine (140 mg, 0.37 mmol) and N-(phenoxycarbonyl) hydroxylamine (112 mg, 0.74 mmol) in anhydrous THF (4 mL) was stirred for 16 hours at 70° C. The solution was then cooled and concentrated under reduced pressure and dried in vacuo. The crude material was purified by column chromatography with silica gel (50:1:0.1 CH2Cl2/MeOH/NH4OH) to give COMPOUND 308 as a white solid (82 mg, 51%).

1. *MAKESOLUTION with 1 (140 mg, 0.37 mmol) and 2 (112 mg, 0.74 mmol) and 3 (4 mL);*
2. *STIR for 16 hours at 70°C;*
3. *CONCENTRATE;*
4. *DRYSOLID under vacuum;*
5. *YIELD -1 (82 mg, 51%).*

Figure 2: Illustration of the experimental procedure prediction task and its dataset curation process. We employ the actions defined by (Vaucher et al., 2021) and the description to action model from (Christofidellis et al., 2023).

step-by-step actions for experimental execution through interpreting chemical reactions (Vaucher et al., 2021), which has a significant value for automating chemical synthesis processes (Vaucher et al., 2020; Zeng et al., 2023). This task aligns well with our focus on RTM, requiring an understanding of chemical reactions and a textual interface to articulate experimental steps. Unfortunately, the absence of public datasets hinders further research and development in this area.

Addressing the identified research gaps, we propose Reaction-Contextualized Molecule-Text Pretraining (**ReactXT**), aiming to improve the text-based understanding of chemical reactions and molecules. Further, we construct an open-source dataset for experimental procedure prediction (**OpenExp**), serving as a key benchmark to evaluate RTM methods. Below, we elaborate on their details.

ReactXT aims to improve the learning paradigm of RTM by introducing three types of input contexts, each of which corresponds to a pretraining task to improve LMs' understanding of chemical reactions or individual molecules. As Figure 1d depicts, the forward reaction context is crafted to learn the chemical connections among molecules involved in the same reaction. These connections are grounded on chemical reaction principles, such as the conservation laws (Atkins and Jones, 2007). Building on this molecular interplay, we hypothesize that understanding other molecules in the same reaction and their descriptions can help predict the current molecule and its textual description. ReactXT encourages LMs to harness these inter-molecule relationships to improve their ability to generate molecular descriptions in reactions and, in turn, deepen their understanding of chemical reaction principles. Further, a backward reaction context is introduced to support retrosynthesis tasks (*cf.* Section 3.1). Finally, as Figure 1c illustrates, ReactXT includes the random molecule context, cultivating the LMs' understanding of individual molecules outside their reactions.

OpenExp features 274, 439 pairs of chemical reactions and their corresponding step-by-step instructions of experimental procedures. This dataset, compiled from the USPTO-Applications (Lowe,

2017) and ORD (Kearnes et al., 2021) databases, will be released under the CC-BY-SA license. To ensure data quality, we have conducted careful data preprocessing. Further, we invite human experts to evaluate the dataset quality. Out of 100 randomly chosen samples, 50 samples could be directly used without any human intervention, and 90 samples required only minor modifications for experimental execution (*cf.* Figure 5).

Our contributions can be summarized as follows:

- We propose ReactXT, a method that incorporates three types of input contexts to incrementally pretrain an LM. These contexts are tailored to enhance LMs' understanding of chemical reactions and individual molecules.

- We curate an open-source experimental procedure prediction dataset OpenExp, a new benchmark for automating chemical synthesis research.

- ReactXT achieves state-of-the-art performances for experimental procedure prediction on the OpenExp dataset, highlighting its superior RTM ability. It also outperforms baselines by 3.2% for molecule captioning on the PubChem324k dataset. ReactXT has competitive performances for retrosynthesis, and we are refining it to surpass the current state-of-the-art method.

2 Related Works

Molecule-Text Modeling (MTM). MTM aims to jointly model molecules and texts to address text-related molecular tasks (Edwards et al., 2022, 2021). Molecules can be represented by 1D sequences of SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2020), making it feasible to pretrain unified LMs on mixed 1D sequences of texts and molecules (Taylor et al., 2022; Edwards et al., 2022; Chithrananda et al., 2020; Zeng et al., 2022). Further, these LMs can be aligned to human preference via instruction tuning (Christofidellis et al., 2023; Fang et al., 2023). In parallel to 1D LMs, multi-modal methods are also studied, using graph neural networks (GNNs) (Hu et al., 2020; Liu et al., 2023c) to encode 2D molecular graphs. Notably, CLIP-style (Radford et al., 2021) cross-modal contrastive learning and BLIP2-style (Li et al., 2023) cross-modal projector are both investigated to facilitate molecule-text retrieval (Su et al., 2022; Liu et al., 2022), and molecule-to-text generation (Liu et al., 2023b; Li et al., 2024), respectively. Recently, MolTC (Fang et al., 2024) is also proposed to model molecular interactions using chain of thoughts. However, prior works mainly focus on individual molecules rather than chemical reactions. To bridge the gap, ReactXT explores reaction-text modeling, facilitating reaction-relevant tasks with a text interface and textual knowledge.

Experimental Procedure Prediction. Synthesizing complex compounds requires detailed planning of synthetic pathways and intermediate steps, a process that is both labor-intensive and complex. Machine learning (ML) can potentially automate the process by predicting experimental procedures. Prior works have explored predicting reaction conditions (*e.g.,* catalyst and solvent) (Gao et al., 2018) and sequences of synthesis steps (Vaucher et al., 2021) by reading chemical reactions. Given known experimental procedures, ML is also explored to empower chemical lab robots (Burger et al., 2020), and automated lab pipelines (Coley et al., 2019; Nicolaou et al., 2020). Notably, tool-augmented GPT4 (OpenAI, 2023) is explored to plan and execute known chemical experiments (Boiko et al., 2023). Unlike prior works, our OpenExp dataset is the first open-source dataset to facilitate the procedure prediction of unseen chemical experiments.

Retrosynthesis and Chemical Reaction Prediction. Given a chemical reaction, retrosynthesis is to predict reactants from products and reaction prediction is to predict products from reactants (Schwaller et al., 2022). They can be formalized as sequence-to-sequence translation represented by SMILES strings (Liu et al., 2017; Irwin et al., 2022; Zhong et al., 2022; Tetko et al., 2020; Ucak et al., 2022). Concurrently, 2D molecular graphs are explored for reaction prediction: selection-based methods focus on classifying the most suitable reaction templates (Chen and Jung, 2021; Dai et al., 2019); and graph-based generative models directly synthesize target molecules (Shi et al., 2020; Sacha et al., 2021; Yan et al., 2020). However, the methods above leverage only reactions without texts. While notably two pioneer works apply ChatGPT for reaction prediction (Shi et al., 2023; Bran et al., 2023), their performances are limited to exploring only prompt engineering.

3 ReactXT: Reaction-Contextualized Molecule-Text Pretraining

ReactXT consists of two key components: 1) the method of creating input contexts to incrementally

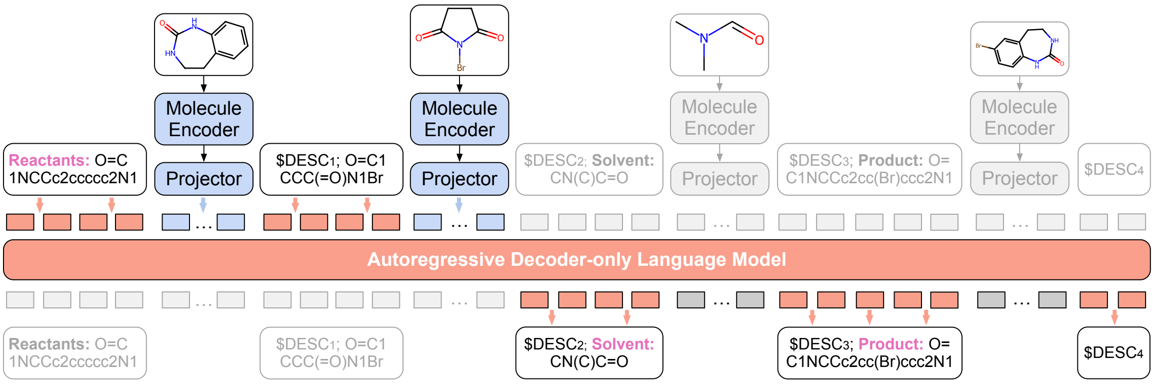

Molecule Encoder
Projector
Molecule Encoder
Projector
Molecule Encoder
Projector
Molecule Encoder
Projector
Reactants: O=C 1NCCc2ccccc2N1
$DESC1; O=C1 CCC(=O)N1Br
$DESC2; Solvent: CN(C)C=O
$DESC3; Product: O= C1NCCc2cc(Br)ccc2N1
$DESC4
Autoregressive Decoder-only Language Model
Reactants: O=C 1NCCc2ccccc2N1
$DESC1; O=C1 CCC(=O)N1Br
$DESC2; Solvent: CN(C)C=O
$DESC3; Product: O= C1NCCc2cc(Br)ccc2N1
$DESC4

Figure 3: Illustration of Reaction-Contextualized Molecule-Text Pretraining. Example uses forward reaction context.

Context Type	Prompt Template
Forward reaction	$\texttt{Reactants: } \underbrace{\texttt{\$SMI}_1\ \texttt{<Mol}_1\texttt{>}\ \texttt{\$DESC}_1\texttt{;}}_{\times n\texttt{: Number of reactants}}\ \texttt{Solvent: \$SMI}_{n+1}\ \texttt{<Mol}_{n+1}\texttt{>}\ \texttt{\$DESC}_{n+1}\texttt{; Product: \$SMI}_{n+2}\ \texttt{<Mol}_{n+2}\texttt{>}\ \texttt{\$DESC}_{n+2}\texttt{<STOP>}$
Backward reaction	$\texttt{Product: \$SMI}_1\ \texttt{<Mol}_1\texttt{>}\ \texttt{\$DESC}_1\texttt{; Solvent: \$SMI}_2\ \texttt{<Mol}_2\texttt{>}\ \texttt{\$DESC}_2\texttt{; Reactants: } \underbrace{\texttt{\$SMI}_3\ \texttt{<Mol}_3\texttt{>}\ \texttt{\$DESC}_3}_{\times n\texttt{: Number of reactants}}\texttt{<STOP>}$
Random molecule	$\texttt{\$SMI}_1\ \texttt{<Mol}_1\texttt{>}\ \texttt{\$DESC}_1\texttt{; \$SMI}_2\ \texttt{<Mol}_2\texttt{>}\ \texttt{\$DESC}_2\texttt{; \$SMI}_3\ \texttt{<Mol}_3\texttt{>}\ \texttt{\$DESC}_3\texttt{; \$SMI}_4\ \texttt{<Mol}_4\texttt{>}\ \texttt{\$DESC}_4\texttt{<STOP>}$

Table 1: Prompt templates for creating input contexts. $\texttt{<Mol}_\texttt{i}\texttt{>}$ is the placeholder for the 2D graph embedding of the i-th molecule; $\texttt{\$SMI}_\texttt{i}$ and $\texttt{\$DESC}_\texttt{i}$ is the SMILES and textual description for the i-th molecule, respectively.

[Abstract] The invention relates to indole acetic acid compounds which function as antagonists of the CRTH2 receptor. The invention also relates to the use of these compounds to inhibit the binding of prostaglandin D2 and its metabolites or certain thromboxane metabolites to the CRTH2 receptor and to treat disorders responsive to such inhibition. **[Properties]** Molecular Weight: 547.60; XLogP3: 6.10; Hydrogen Bond Donor Count: 0; Hydrogen Bond Acceptor Count: 7; Rotatable Bond Count: 8; Exact Mass: 547.19; Monoisotopic Mass: 547.19; Topological Polar Surface Area: 89.40; Heavy Atom Count: 39; Formal Charge: 0; Complexity: 1020; Isotope Atom Count: 0; Defined Atom Stereocenter Count: 0; Undefined Atom Stereocenter Count: 0; Defined Bond Stereocenter Count: 0; Undefined Bond Stereocenter Count: 0; Covalently-Bonded Unit Count: 1; Compound Is Canonicalized: Yes.

Table 2: Molecule description example, including the patent abstract and the computed/experimental properties. The described molecule is Cc1c(C2=NN(CCc3ccccc3)S(=O)(=O)c3ccccc32)c2cc(F)ccc2n1CC(=O)OC(C)(C)C.

pretrain an LM, and 2) a balanced sampling strategy for the reaction contexts. We begin by introducing our multi-modal LM backbone, then proceed to elaborate on ReactXT's two components.

Multi-Modal Language Model Backbone. Molecules can be represented by their 1D SMILES or 2D molecular graphs (Wells, 2012). We employ MolCA (Liu et al., 2023b) as our primary LM backbone to effectively harness both the 1D and 2D molecular modalities. Specifically, MolCA incorporates a GNN encoder (You et al., 2020) for encoding 2D molecular graphs. This GNN's output then is mapped to an LM's (*i.e.*, Galactica; Taylor et al. (2022)) input space via a cross-modal projector, thereby enabling the LM to perceive 2D molecular graphs. Both the cross-modal projector and the GNN have been pretrained for molecule-text alignment (Li et al., 2023). MolCA shows promising performances when finetuned for molecule captioning and IUPAC name prediction.

3.1 Creating Input Contexts

Addressing the core challenges of LMs hinges on the careful selection of the input data. As shown in Table 1, ReactXT incorporates three types of input contexts to incrementally pretrain LMs: forward reaction context, backward reaction context, and random molecule context. These contexts are tailored for a text-based understanding of chemical reactions and individual molecules:

- **Forward Reaction Context.** As Figure 3 illustrates, the forward reaction context labels molecules according to their roles – `Reactant`, `Catalyst`, `Solvent`, and `Product` – in the reaction, and arranges them in this specific sequential order. Note, not every reaction has a `Catalyst` or `Solvent`. For each molecule, we append its 2D molecular graph embeddings (*e.g.*, $\texttt{<Mol}_1\texttt{>}$; Liu et al. (2023b)) after its SMILES to enhance the LM's understanding of molecular structures; and append molecular descriptions (*e.g.*, $\texttt{\$DESC}_1$) following the 2D molecular graph embeddings to align molecules with texts.

- **Backward Reaction Context.** Similar to the forward context but with the order of molecular roles reversed, this context aims to combat the Reversal Curse (Berglund et al., 2023) of LMs:

LMs trained on "A is B" fail to generalize to "B is A". The reversal generalization is crucial because downstream applications include backward retrosynthesis (Schwaller et al., 2022).

- **Random Molecule Context.** Introduced to ensure LMs retain the capability to describe individual molecules outside chemical reactions.

Context Length. In each input context, we use up to k molecules and their descriptions, where k is a hyperparameter. For reactions with over k molecules, we apply weighted molecule sampling, as explained in Section 3.2.

Molecule Descriptions. One crucial component of the input contexts is the molecule description, whose quality and comprehensiveness are vital for molecule-text alignment. We collect molecular descriptions and properties from multiple sources, encompassing three types of content:

- **Molecule Patent Abstracts.** We source patent abstracts from PubChem's Patent View[*]. These abstracts typically describe molecular structures, properties, or applications, but may also include irrelevant information if the molecule is merely mentioned in passing rather than being the central subject. Despite the noise, patent abstracts are indispensable for RTM: they cover $\sim$95% molecules in our employed reaction databases (Lowe, 2017; Kearnes et al., 2021). In contrast, the molecule-text datasets (Liu et al., 2022, 2023b) derived from PubChem's description section only cover $\sim$1% of these molecules.
- **Computed and Experimental Properties.** We retrieve these numerical properties from PubChem, aiming to enhance the understanding of molecular structures through predictive learning. Certain properties are also helpful for reaction prediction. For example, knowing the solubility helps determine concentrations when preparing solutions; the knowledge of melting and boiling points helps identify the states of matter at given temperatures. Table 2 shows an example of a patent abstract and computed/experimental properties. Table 14 includes detailed statistics of our collected molecule properties.
- **PubChem Descriptions.** Following (Liu et al., 2022, 2023b), we employ molecular descriptions from PubChem. Due to their limited coverage ($\sim$1%) for molecules in reaction databases (Lowe, 2017; Kearnes et al., 2021), we incorporate them exclusively for the random molecule context.

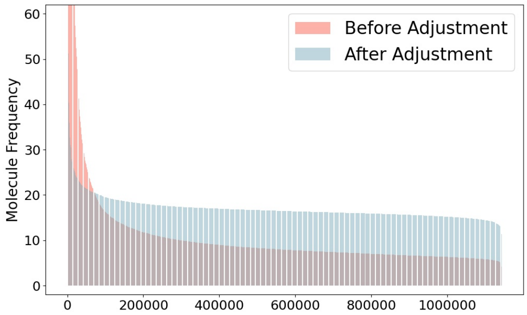


Figure 4: Distribution of molecules in the pretraining chemical reactions. For after adjustment, we conduct weighted sampling of chemical reactions matching the size of the pretraining dataset.

Autoregressive Language Modeling for Interleaved Molecule-Text Sequences. Given the input contexts above of interleaved molecules and texts, we apply language modeling loss to incrementally pretrain the LM, molecule encoder, and projector. We compute loss only for text tokens, excluding 2D molecular graph embeddings.

3.2 Balanced Sampling of Reaction Contexts

Figure 4 reveals a skewed distribution of molecules in chemical reactions (the red bars), with a small group of molecules appearing far more frequently than others. To address this imbalance, we develop a sampling strategy that promotes a fairer representation of molecules across reactions. This method reduces the dominance of commonly occurring molecules by adjusting 1) the sampling weight of each reaction r: $W(r)$, and 2) the sampling weight of each molecule m within a chosen reaction r: $W(m|r)$, based on the equations below:

$$W(r) = \frac{\sum_{m \in r} 1/\text{Count}(m)}{\sum_{r' \in \mathcal{R}} \sum_{m \in r} 1/\text{Count}(m)}, \quad (1)$$

$$W(m|r) = \frac{1/\text{Count}(m)}{\sum_{m' \in r} 1/\text{Count}(m')}, \quad (2)$$

where $\mathcal{R}$ denotes the dataset of chemical reactions; Count(m) denotes molecule m's count in $\mathcal{R}$.

Equation (1) sets a reaction's sampling weight inversely to the total occurrences of its molecules, favoring reactions with rare molecules; Equation (2) boosts the weights of rarer molecules within a

[*] https://pubchem.ncbi.nlm.nih.gov/docs/patents

Total reactions	2262637	100%
Too large perplexity score	329160	14.55%
More than one product	105577	4.67%
Incomplete mapping of molecules (from chemical equation)	1034908	45.74%
Incomplete mapping of molecules (from action sequence)	178689	7.90%
Remove duplicate reactions	254099	11.23%
Filter out too short actions	14022	0.62%
Other errors	71743	3.16%
Remaining reactions	274439	12.13%

Table 3: Preprocessing steps and the number of samples removed at each step.

Dataset	Total	Train	Valid	Test	Open Source
Vaucher et al. (2021)	693k	555k	69k	69k	No
OpenExp, Ours	274k	220k	27k	27k	Yes

Table 4: Dataset statistics and comparison to prior work.

given reaction. These weights are then applied for weighted random sampling without replacement (Efraimidis and Spirakis, 2006). The blue bars in Figure 4 present the sampling frequency of molecules after adjustment, showing a flatter distribution. Implementation details are in Appendix B.

4 OpenExp: An Open-Source Dataset for Experimental Procedure Prediction

Here we briefly introduce OpenExp's curation process and defer the details to Appendix A.1. OpenExp is sourced from chemical reaction databases of USPTO-Applications (Lowe, 2017) and ORD (Kearnes et al., 2021). As illustrated in Figure 2, these databases include chemical reactions and the corresponding unstructured descriptions of experimental procedures. To convert these unstructured descriptions into structured action sequences, we first run the pragraph2action model from (Christofidellis et al., 2023), and then conduct preprocessing following (Vaucher et al., 2021). The preprocessing is to remove low-quality data, eliminate duplicates, and construct molecule mapping between reactions and experimental procedures. Specific preprocessing steps are summarized in Table 3. An example is shown in Table 11.

As shown in Table 4, the final OpenExp dataset includes 274k reaction-procedure pairs. It is randomly divided into train/valid/test sets by the 8:1:1 ratio. Compared to the prior work (Vaucher et al., 2021), which is closed-source for using the commercial Pistachio database[†], we open-source this dataset to assist future research.

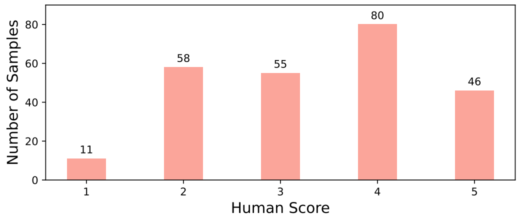


Figure 5: Human evaluations on OpenExp.

To obtain insights on dataset quality, we invite two graduate students in chemistry to rate the alignment between the action sequences and their original descriptions, on a scale from 1 (lowest) to 5 (highest), as depicted in Figure 5. Briefly, of the total 250 samples evaluated, 126 ($\geq 50\%$) action sequences have at most 1 error (scores above 4), and 181 ($\geq 50\%$) action sequences have at most 2 errors (scores above 3). Our closer inspection shows that the one error in score-4 samples is usually a typo of material/action name, or a discrepancy of numerical value, and does not impede the overall execution. See Appendix C.3.2 for details.

5 Experiment

We empirically evaluate ReactXT across three downstream tasks, including experimental procedural prediction, molecule captioning, and retrosynthesis. Further, we include ablation studies showcasing the contributions of individual components. To ensure the significance of our experimental, we include statistical tests results in Appendix C.2.

5.1 Experimental Setting

ReactXT is initialized by the stage-2 checkpoint of MolCA$_{1.3B}$ (Liu et al., 2023b), if not specially noted. It is then pretrained using our proposed method, and subsequently finetuned for each downstream dataset separately. The context length k is 4. We employ full-parameter tuning for pretraining and finetuning. More details are in Appendix B.

ReactXT's Pretraining Dataset. Our pretrain dataset includes PubChem324k's pretrain subset (Liu et al., 2023b), which includes 298k molecule-text pairs, and 1.11 million chemical reactions from the USPTO-Applications (Lowe, 2017) and ORD (Kearnes et al., 2021) databases. For molecules in reactions, we obtain their patent abstracts and molecular properties following Section 3.1. To prevent information leakage, we have

[†] https://www.nextmovesoftware.com/pistachio

Method	Validity	BLEU-2	BLEU-4	100%LEV	90%LEV	75%LEV	50%LEV	ROUGE-1	ROUGE-2	ROUGE-L
Random, among all reactions	63.2	34.5	19.1	0.0	0.0	0.0	13.6	46.6	18.1	36.4
Random, compatible pattern	**100.0**	37.8	22.1	0.0	0.0	0.1	16.5	47.8	21.0	38.4
Nearest neighbor	76.0	45.0	30.7	0.6	6.5	13.0	38.4	55.7	29.2	47.0
TextChemT5$_{220M}$	99.3	54.1	40.6	0.4	4.6	13.7	61.2	61.5	40.3	56.4
MolT5-Large$_{780M}$	99.6	54.5	41.0	0.6	6.6	16.6	63.7	62.5	40.9	57.2
Galactica$_{1.3B}$	99.9	53.5	39.5	0.4	5.7	13.4	60.5	60.9	38.6	55.2
MolCA, Galac$_{1.3B}$	99.9	54.9	41.5	**1.0**	9.2	18.9	65.3	62.5	40.4	57.0
ReactXT, Galac$_{1.3B}$, Ours	**100.0**	**57.4**	**44.0**	**1.0**	**9.5**	**22.6**	**70.2**	**64.4**	**42.7**	**58.9**

Table 5: Comparison of experimental procedure prediction performances (%) on the OpenExp dataset. The subscript denotes each model's parameter size. We conduct full-parameter fine-tuning for all models.

Method	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small$_{80M}$	14.8	8.5	26.5	13.5	23.6	18.5
MolT5-Base$_{250M}$	30.1	20.9	40.3	25.1	33.8	35.6
MolT5-Large$_{780M}$	30.2	22.2	41.5	25.9	34.8	36.6
Galactica$_{1.3B}$, LoRA ft	34.6	26.9	46.3	32.3	41.5	41.1
MoMu-Small$_{82M}$	19.1	12.0	29.7	16.3	26.7	21.8
MoMu-Base$_{252M}$	30.2	21.5	40.5	25.1	34.4	34.2
MoMu-Large$_{782M}$	31.1	22.8	41.8	25.7	36.7	36.2
MolCA, MolT5-Large$_{877M}$	32.9	26.3	49.8	35.7	44.2	42.4
MolCA, Galac$_{125M}$	31.9	24.3	47.3	33.9	43.2	41.6
MolCA, Galac$_{1.3B}$, LoRA ft	38.7	30.3	50.2	35.9	44.5	45.6
MolCA, Galac$_{1.3B}$, full ft*	39.4	32.2	52.7	39.4	47.6	49.2
ReactXT, Galac$_{1.3B}$, Ours	**42.6**	**35.2**	**54.7**	**41.7**	**49.6**	**51.2**

(a) PubChem324k dataset.

Method	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
MolT5-Small$_{80M}$	51.9	43.6	62.0	46.9	56.3	55.1
MolT5-Base$_{250M}$	54.0	45.7	63.4	48.5	57.8	56.9
MolT5-Large$_{780M}$	59.4	50.8	65.4	51.0	59.4	61.4
TextChemT5$_{60M}$	56.0	47.0	63.8	48.8	58.0	58.8
TextChemT5$_{220M}$	62.5	54.2	68.2	54.3	62.2	64.8
MoMu-Small$_{82M}$	53.2	44.5	-	-	56.4	55.7
MoMu-Base$_{252M}$	54.9	46.2	-	-	57.5	57.6
MoMu-Large$_{782M}$	59.9	51.5	-	-	59.3	59.7
MolCA, Galac$_{125M}$	61.2	52.6	67.4	52.1	60.6	63.6
MolCA, Galac$_{1.3B}$, LoRA ft	62.0	53.1	68.1	53.7	61.8	65.1
ReactXT, Galac$_{1.3B}$	**62.9**	**55.0**	**69.2**	**56.0**	**63.4**	**66.4**

(b) CheBI-20 dataset.

Table 6: Molecule captioning performance (%) on the PubChem324k and CheBI-20 datasets. * denotes our re-implementation. Other baseline results are borrowed from (Liu et al., 2023b; Christofidellis et al., 2023).

Method	Top-1	Top-3	Top-5	Top-10
MEGAN	48.1	70.7	78.4	86.1
AT	53.5	-	81.0	85.7
Chemformer	54.3	-	62.3	63.0
Train with aug., test without aug.				
R-SMILES	51.2	**74.9**	**81.1**	**83.0**
MolT5-Large$_{780M}$*	<u>53.9</u>	69.9	74.6	77.3
ReactXT, Galac$_{1.3B}$, Ours	**54.4**	<u>73.6</u>	<u>78.9</u>	**83.0**
Train with aug., test with aug.				
R-SMILES	<u>56.3</u>	<u>79.2</u>	<u>86.2</u>	**91.0**
MolT5-Large$_{780M}$*	56.0	76.0	80.7	85.1
ReactXT, Galac$_{1.3B}$, Ours	**58.6**	**81.1**	**86.5**	**91.0**

Table 7: Retrosynthesis accuracies (%) on USPTO-50K. * denotes our re-implementation. Other baselines are from (Zhong et al., 2022). In each part, **bold** denotes the best result, and <u>underline</u> denotes the second best.

excluded 54k reactions that appear in the valid/test sets of the downstream datasets (*i.e.,* OpenExp, USPTO-50K (Schneider et al., 2016)) from the initial collection of 1.16 million reactions. See Appendix A.2 for more details.

Baselines. We compare ReactXT with the state-of-the-art LMs in science domain, including Galactica (Taylor et al., 2022), MolT5 (Edwards et al., 2022), TextChemT5 (Christofidellis et al., 2023), and MolCA (Liu et al., 2023b). For retrosynthesis and forward reaction prediction tasks, we also compare with task-specific LMs: R-SMILES (Zhong et al., 2022), AT (Tetko et al., 2020), MEGAN (Sacha et al., 2021), and Chemformer (Irwin et al., 2022). For captioning, we additionally compare against MoMu (Su et al., 2022).

Pretrain Input Context	Pretrain Data Type	BLEU-2	BLEU-4	75%LEV	50%LEV	ROUGE-1	ROUGE-2	ROUGE-L
No incremental pretrain	-	54.9	41.5	18.9	65.3	62.5	40.4	57.0
Random molecules	reaction, sing. mol.	56.6	43.2	20.9	69.4	63.8	41.9	58.3
Reactions w/o bal. samp.	reaction	56.8	43.3	21.3	69.2	64.0	42.1	58.5
Reactions	reaction	57.1	43.8	22.2	70.1	64.3	42.6	**58.9**
ReactXT	reaction, sing. mol.	**57.4**	**44.0**	**22.6**	**70.2**	**64.4**	**42.7**	**58.9**

Table 8: Ablation study of input contexts for incrementally pretrain MolCA, Galac$_{1.3B}$. Results are for experimental procedure prediction. Reactions denote both the forward reaction context and the backward reaction context.

Backbone LM	BLEU-2	BLEU-4	75%LEV	50%LEV	ROUGE-1	ROUGE-2	ROUGE-L
MolT5-Large$_{780M}$	54.5	41.0	16.6	63.7	62.5	40.9	57.2
MolT5-Large$_{780M}$, ReactXT pretrain	**55.6**	**42.1**	**17.2**	**66.6**	**63.6**	**41.7**	**58.1**
Galactica$_{1.3B}$	53.5	39.5	13.4	60.5	60.9	38.6	55.2
Galactica$_{1.3B}$, ReactXT pretrain	**56.5**	**43.1**	**20.8**	**68.7**	**63.7**	**41.8**	**58.2**
MolCA, Galac$_{1.3B}$	54.9	41.5	18.9	65.3	62.5	40.4	57.0
MolCA, Galac$_{1.3B}$, ReactXT pretrain	**57.1**	**43.8**	**22.2**	**70.1**	**64.3**	**42.6**	**58.9**

Table 9: Ablation study of ReactXT pretraining for experimental procedure prediction.

5.2 Experimental Procedure Prediction

Following (Vaucher et al., 2021), we employ the following evaluation metrics: Validity, which checks the syntactical correctness of the action sequence; machine-translation metrics BLUE (Papineni et al., 2002) and ROUGE (Lin, 2004); and the normalized Levenshtein similarity (Levenshtein et al., 1966). Specifically, 90%LEV denotes the proportion of predictions with a normalized Levenshtein score larger than 0.9. The three naive baselines based on random sampling and nearest neighbor are borrowed from (Vaucher et al., 2021). See Appendix B for details.

Table 5 presents the performances. We can observe that ReactXT consistently outperforms baselines across all metrics. Specifically, it surpasses baselines by 2.2% for BLEU-2 and 3.3% for 75%LEV, demonstrating ReactXT's effectiveness for text-based reaction understanding.

5.3 Molecule Captioning

To evaluate ReactXT's ability to understand single-molecules, we present its performances of molecule captioning on the PubChem324k (Liu et al., 2023b) and CheBI-20 (Edwards et al., 2022) datasets. We report metrics of BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005).

Table 6 presents the captioning performances. We can observe that ReactXT consistently outperforms the baselines. Specifically, ReactXT shows improvements of 3.2% BLEU-2 and 2.3% ROUGE-2 scores on PubChem324k, and 1.7% ROUGE-2 on CheBI-20. These improvements underscore the effectiveness of our pretraining method for enhancing understanding of individual molecules.

5.4 Retrosynthesis

Retrosynthesis is to predict the reactant molecules given the product molecules. For this task, we employ the evaluation metrics of top-k accuracy, which measures the percentage of exact match to the ground truth in the top-k predictions. Following (Zhong et al., 2022), we conduct self-supervised pretraining on the USPTO-full(Dai et al., 2019) dataset and use the root-aligned augmentations of SMILES during training and testing. Additionally, we report performances of testing without these augmentations.

Table 7 presents the results. ReactXT outperforms R-SMILES across all metrics when testing with augmentations. Notably, the improvement in top-1 accuracy is particularly significant, achieving a 2.3% increase over the second best value. Regardless of whether test set data augmentation is applied, ReactXT achieves better top-k accuracies than MolT5-Large, which is also a multimodal LM. These performance improvements stem from ReactXT's use of reactions for pretraining, rather than individual molecules.

5.5 Ablation Study

In this section, we conduct ablation studies to show the impact of different pretrain data types and backbone LMs in our method.

Pretrain Data Type. We ablate the key components of ReactXT, using the baseline of MolCA, Galac$_{1.3B}$ without incremental pretraining. Table 8 presents the results. Specifically, we compare three variants of ReactXT: 1) pretraining with solely the random molecule contexts using the same pretrain dataset; 2) pretraining with forward and backward reaction contexts without the random molecule context; and 3) applying uniform sampling on reaction contexts instead of balanced sampling.

We can observe that 1) ReactXT's full model shows the best performance, showing its performance is the integrated contribution of all components; 2) applying random molecule contexts alone improves upon the baseline, underscoring the valuable textual knowledge from our meticulously crafted pretraining dataset; 3) incorporating reaction contexts yields better results than random molecule contexts, highlighting the benefits of learning reaction knowledge during pretraining; and 4) balanced sampling improves the performance upon uniform sampling.

Backbone LMs. We conduct ablation studies on the backbone LMs. This study involves three different molecular-text LMs: 1) MolCA, which represents molecules using both 1D SMILES and 2D graphs, based on a decoder-only architecture; 2) Galactica, which represents molecules using 1D SMILES, based on a decoder-only architecture; and 3) MolT5, which represents molecules using 1D SMILES, based on an encoder-decoder architecture. The experimental results are presented in Table 9. We can observe that the ReactXT pretraining scheme achieves consistent performance improvements, regardless of the backbone language model used.

6 Conclusion and Future Works

In this work, we explore reaction-text modeling to empower reaction-relevant tasks with textual interfaces and knowledge. We present ReactXT, a pretraining method to learn chemical reactions within the context of the corresponding molecular textual descriptions. Additionally, we propose a new dataset OpenExp to support open-source research for experimental procedure prediction. ReactXT establishes the best performances across tasks of experimental procedure prediction and molecule captioning. It presents competitive performances for retrosynthesis.

In future work, we plan to apply LMs to learn the interactions among large molecules (*e.g.,* proteins and nucleic acids), or introduce molecules' dynamics and 3D spatial structures for better molecule-language understanding (Luo et al., 2023).

Limitations

In this and also the previous work (Vaucher et al., 2021), the evaluation for experimental procedure prediction is constrained to the comparison between the predictions and the reference action sequences. While improving this metric does reflect the improvement in experimental design, it should be acknowledged that the evaluation of real-world chemical experiments is preferred for the developed models in future. For this purpose, the methods on automated chemistry pipelines (Boiko et al., 2023; Coley et al., 2019; Nicolaou et al., 2020) can be potentially considered.

Another limitation or future direction is improving the action space defined in our proposed OpenExp dataset, aiming to cover a wider range of chemical experiments. For example, the action of 'Purify' is absent; and the action of 'Concentration' can be refined into operations such as 'Evaporation' and 'Pressurize' for clearer instructions of chemical experiments.

Potential Ethics Impact

In this study, the proposed method and dataset focus on chemical reactions and molecules, and include no human subjects. Consequently, we believe this study presents no direct ethical concerns. However, the inclusion of LMs in our study does raise potential issues, as LMs can be misused to produce incorrect or biased information. Therefore, the ethical implications of our work align with those common to LM research, emphasizing the need for responsible use and application of LMs.

Acknowledgement

This research is supported by the National Science and Technology Major Project (2023ZD0121102), National Natural Science Foundation of China (92270114). This research is partially supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-010-SGIL), the Singapore Ministry of Education Academic Research Fund Tier 1 (Award No: T1 251RES2207) and the Google Cloud Research Credits program with the award (Q4MJ-YH1K-3MVX-FP6Q). This research is supported by NExT Research Center.

References

Peter Atkins and Loretta Jones. 2007. *Chemical principles: The quest for insight*. Macmillan.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, pages 65–72. Association for Computational Linguistics.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. 2023. The reversal curse: Llms trained on" a is b" fail to learn" b is a". *arXiv preprint arXiv:2309.12288*.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Jannis Born and Matteo Manica. 2023. Regression transformer enables concurrent sequence regression and generation for molecular language modelling. *Nat. Mac. Intell.*, 5(4):432–444.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew White, and Philippe Schwaller. 2023. Augmenting large language models with chemistry tools. In *NeurIPS 2023 AI for Science Workshop*.

Benjamin Burger, Phillip M Maffettone, Vladimir V Gusev, Catherine M Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, Ben M Alston, Buyi Li, Rob Clowes, et al. 2020. A mobile robotic chemist. *Nature*, 583(7815):237–241.

Shuan Chen and Yousung Jung. 2021. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. 2023. Unifying molecular and textual representations via multi-task language modelling. In *ICML*.

Connor W Coley, Dale A Thomas III, Justin AM Lummiss, Jonathan N Jaworski, Christopher P Breen, Victor Schultz, Travis Hart, Joshua S Fishman, Luke Rogers, Hanyu Gao, et al. 2019. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365(6453):eaax1566.

Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. 2019. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32.

Carl Edwards, Tuan Manh Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *EMNLP*, pages 375–413. Association for Computational Linguistics.

Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *EMNLP (1)*, pages 595–607. Association for Computational Linguistics.

Pavlos S Efraimidis and Paul G Spirakis. 2006. Weighted random sampling with a reservoir. *Information processing letters*, 97(5):181–185.

Junfeng Fang, Shuai Zhang, Chang Wu, Zhengyi Yang, Zhiyuan Liu, Sihang Li, Kun Wang, Wenjie Du, and Xiang Wang. 2024. MolTC: Towards molecular relational modeling in language models. *arXiv preprint arXiv:2402.03781*.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. 2023. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. *CoRR*, abs/2306.08018.

Hanyu Gao, Thomas J Struble, Connor W Coley, Yuran Wang, William H Green, and Klavs F Jensen. 2018. Using machine learning to predict suitable conditions for organic reactions. *ACS central science*, 4(11):1465–1476.

Taicheng Guo, Kehan Guo, Bozhao Nan, Zhengwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2023. What indeed can GPT models do in chemistry? A comprehensive benchmark on eight tasks. *CoRR*, abs/2305.18365.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. 2020. Strategies for pre-training graph neural networks. In *ICLR*.

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.

Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. 2021. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826.

Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. 2020. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn. Sci. Technol.*, 1(4):45024.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. 2024. 3d-molm: Towards 3d molecule-text interpretation in language models. In *ICLR*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. 2017. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Anima Anandkumar. 2022. Multi-modal molecule structure-text model for text-based retrieval and editing. *CoRR*, abs/2212.10789.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *EMNLP*, pages 15623–15638. Association for Computational Linguistics.

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

Daniel Lowe. 2017. Chemical reactions from US patents (1976-Sep2016).

Yanchen Luo, Sihang Li, Zhiyuan Liu, Jiancan Wu, Zhengyi Yang, Xiangnan He, Xiang Wang, and Qi Tian. 2023. Text-guided diffusion model for 3d molecule generation.

Christos A. Nicolaou, Ian A. Watson, Mark Lemasters, Thierry Masquelin, and Ji-Bo Wang. 2020. Context aware data-driven retrosynthetic analysis. *J. Chem. Inf. Model.*, 60(6):2728–2738.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. Stout: Smiles to iupac names using neural machine translation. *Journal of Cheminformatics*, 13(1):1–14.

Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. 2021. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284.

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

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, Teodoro Laino, and Jean-Louis Reymond. 2019. Data-driven chemical reaction classification, fingerprinting and clustering using attention-based neural networks. *ChemRxiv*.

Philippe Schwaller, Alain C Vaucher, Ruben Laplaza, Charlotte Bunne, Andreas Krause, Clemence Corminboeuf, and Teodoro Laino. 2022. Machine intelligence for chemical reaction space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1604.

Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. 2020. A graph to graphs framework for retrosynthesis prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8818–8827. PMLR.

Yaorui Shi, An Zhang, Enzhi Zhang, Zhiyuan Liu, and Xiang Wang. 2023. ReLM: Leveraging language models for enhanced chemical reaction prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5506–5520. Association for Computational Linguistics.

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. *CoRR*, abs/2209.05481.

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

Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. 2020. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575.

Umit V Ucak, Islambek Ashyrmamatov, Junsu Ko, and Juyong Lee. 2022. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nature communications*, 13(1):1186.

Alain C Vaucher, Philippe Schwaller, Joppe Geluykens, Vishnu H Nair, Anna Iuliano, and Teodoro Laino. 2021. Inferring experimental procedures from text-based representations of chemical reactions. *Nature communications*, 12(1):2573.

Alain C Vaucher, Federico Zipoli, Joppe Geluykens, Vishnu H Nair, Philippe Schwaller, and Teodoro Laino. 2020. Automated extraction of chemical synthesis actions from experimental procedures. *Nature communications*, 11(1):3601.

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

Alexander Frank Wells. 2012. *Structural inorganic chemistry*. Oxford university press.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. 2020. Retroxpert: Decompose retrosynthesis prediction like A chemist. In *NeurIPS*.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *NeurIPS*.

Zheni Zeng, Yi-Chen Nie, Ning Ding, Qian-Jun Ding, Wei-Ting Ye, Cheng Yang, Maosong Sun, E Weinan, Rong Zhu, and Zhiyuan Liu. 2023. Transcription between human-readable synthetic descriptions and machine-executable instructions: an application of the latest pre-training technology. *Chemical Science*, 14(35):9360–9373.

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(1):862.

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

A Dataset Details

A.1 Collection and Preprocessing of OpenExp

OpenExp is compiled from the raw data from the two following sources:

- **USPTO-Applications** (Lowe, 2017). This dataset comprises records of 1.94 million reactions and their corresponding applications from the United States Patent and Trademark Office (USPTO) published between 2001 and September 2016. We download the raw XML files from the Figshare website [‡]. For each reaction in this dataset, we extract its key information from four elements: `<productList>`, which contains the products of the reaction; `<reactantList>`, detailing the reactants; `<spectatorList>`, encompassing the catalysts and solvents; and `<dl:paragraphText>`, which provides a textual description of the experimental procedures.
- **Open Reaction Database** (Kearnes et al., 2021). The ORD [§] dataset contains over 2 million chemical reactions, which include detailed records of reaction conditions and experimental procedures. It includes data from the USPTO applications (2001-2016 Sep), USPTO-granted patents (1976-2016 Sep), and experimental records from chemical literature.

Paragraph2Action. As illustrated in Figure 2, these databases include chemical reactions and the corresponding unstructured descriptions of experimental procedures. The unstructured nature of these descriptions poses a significant challenge to 1) automate chemical synthesis with robots (Vaucher et al., 2020; Burger et al., 2020); and 2) apply ML methods to predict experimental procedures of unseen reactions. To address this, the task of paragraph2action (Vaucher et al., 2020; Zeng et al., 2023) is proposed, aiming to convert unstructured experimental procedure descriptions into structured, step-by-step instructions with pre-defined actions. In this study, we leverage the action space defined by (Vaucher et al., 2020, 2021), and the pragraph2action model released by (Christofidellis et al., 2023).

Preprocessing. Following (Vaucher et al., 2021), we conduct preprocessing after the paragraph2action conversion, The preprocessing has two purposes: 1) extracting the important entities (*i.e.,* molecules) in experimental procedures and mapping all molecules to their precursors in the chemical reaction; 2) applying a rule-based filtration to improve the dataset quality. Our preprocessing strategy is inspired by (Vaucher et al., 2020), augmented with additional 2 steps: perplexity filtering and similar action aggregation. The complete preprocessing steps are listed below:

Action	Occurrence	Action	Occurrence
`Add`	744,533	`Wait`	38,211
`Stir`	287,413	`Recrystal.`	25,600
`Concentrate`	276,551	`PhaseSepa.`	24,141
`Yield`	274,439	`PH`	21,756
`MakeSolution`	272,537	`Quench`	18,699
`Filter`	247,625	`Partition`	16,045
`Wash`	224,286	`Triturate`	13,390
`DrySolution`	178,248	`DrySolid`	6,435
`CollectLayer`	146,379	`Degas`	4,789
`Extract`	114,855	`Microwave`	2,237
`SetTemp.`	44,126	`Sonicate`	450
`Reflux`	43,296		

Table 10: Action space and actions' occurrences in the OpenExp dataset.

- Perplexity Filtering. To ensure the quality of the above translation step, we compute a perplexity score for each output and exclude samples with a score larger than 1.0. These perplexity scores are calculated using the TextChemT5 model.
- Entity Recognition. We extract all the molecules (either by name or SMILES) from the action sequences using the source codes of (Vaucher et al., 2020). Then, we conduct string matching of IUPAC names between the extracted molecules and those in the chemical reactions. STOUT (Rajan et al., 2021) and PubChemPy[¶] are used for the translation between IUPAC names and SMILES. If any molecule cannot be matched with its counterpart in the chemical reactions, we consider the reaction data invalid and remove it from the dataset. However, we permit the inclusion of certain common substances, such as common organic solvents, in every reaction. The names and SMILES expressions of the 134 common substances are included in our code. After entity recognition, we assign each entity a unique ID and update the experimental procedures by replacing the entity mentions with the corresponding entity IDs.

[‡] https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873?file=8664370

[§] https://open-reaction-database.org

[¶] https://github.com/mcs07/PubChemPy

Field	Value
Reactant	$1$: OC(CCc1ccccn1)C(F)(F)F $3$: CC(C)(C)[Si](C)(C)Cl $4$: c1c[nH]cn1
Solvent	$2$: ClCCl
Catalyst	$5$: CN(C)c1ccncc1
Product	$-1$: CC(C)(C)[Si](C)(C)OC(CCc1ccccn1)C(F)(F)F
Experimental Procedures	**MAKESOLUTION** with $1$ and $2$ (10 mL) ; **ADD** $3$ (616 mg, 4.1 mmol, 1.2 eq) at 0°C ; **ADD** $4$ (697 mg, 10.2 mmol, 3.0 eq) at 0°C ; **ADD** $5$ (415 ng, 3.4 mmol) at 0°C ; **STIR** for 36 hours ; **CONCENTRATE** ; **YIELD** $-1$ (970 mg, 89%).
Source	A solution of 700 mg (3.4 mmol) of 1,1,1-trifluoro-4-pyridin-2-ylbutan-2-ol in 10 mL of dichloromethane was treated with 616 mg (4.1 mmol, 1.2 eq.) of tert-butyldimethylsilyl chloride, 697 mg (10.2 mmol, 3.0 eq.) of imidazole and 415 ng (3.4 mmol) of 4-dimethylaminopyridine at 0° C. The resulting mixture was allowed to warm to room temperature and as stirred for 36 hours. Then the mixture w was concentrated and the residue was purified by flash chromatography to give 970 mg (89%) of 2-[3-(tert-butyldimethylsilanyloxy)-4,4,4-trifluorobutyl]pyridine as a colorless oil.

Table 11: Illustrative example of the OpenExp dataset. **BOLDED BLUE** indicates pre-defined action.

Total reactions	2262637	100%
Too large perplexity score	329160	14.55%
More than one product	105577	4.67%
Incomplete mapping of molecules (from chemical reaction)	1034908	45.74%
Incomplete mapping of molecules (from action sequence)	178689	7.90%
Remove duplicate reactions	254099	11.23%
Filter out too short actions	14022	0.62%
Other errors	71743	3.16%
Remaining reactions	274439	12.13%

Table 12: Number of samples removed at each preprocessing step.

- Common Substance Renaming. We standardized the nomenclature for common substances that are known by multiple names (*e.g.,* water may also be referred to as H2O, pure water, water (aq.), *etc.*) to improve the dataset's precision. Using PubChemPy, we align the different names to their standardized SMILES representations, allowing us to identify when different terms refer to the same molecule by comparing their SMILES expressions.

- Similar Action Aggregation. If two adjacent operations are highly similar (*e.g., STIR* and *STIR for 5 min*), they are merged together.

- Ensuring Single Product. This dataset focuses on the preparation of a single material, hence we remove reactions that yield multiple products.

- Action Filtering. We remove action sequences that have fewer than five actions or contain invalid actions.

- Reaction Deduplication. We remove the duplicated reactions from the dataset.

Table 12 presents the number of samples removed at each preprocessing step. Further, Table 11 provides an example from the final OpenExp dataset, we can observe that it encompasses:

- Structured, step-by-step instructions of experimental procedures;

- All molecules in the reaction and their roles (*i.e.,* reactant, solvent, catalyst, product).

- The mapping between the recognized entities (*i.e.,* molecules) and their IDs.

- The original unstructured experimental procedures.

Discussion on License. The ORD database is accessible under the CC-BY-SA license, and the USPTO-Applications dataset is available under the CC0 license. We have used codes from TextChemT5 (Christofidellis et al., 2023) and Paragraph2Actions (Vaucher et al., 2021), which are both licensed under the MIT license. Therefore, we will release OpenExp under the CC-BY-SA license to comply with the most restrictive license of these resources. This license permits content distribution and sharing, provided the same license is applied.

Human Evaluation. We invite two PhD students majoring in chemistry to evaluate the quality of the OpenExp dataset. Specifically, 250 data points are randomly sampled from the dataset, and assigned to the evaluators according to the following rules: 1) the first 50 data points are assigned to both volunteers simultaneously to verify the consistency of their evaluations; 2) the remaining 200 data points are then evenly assigned to the two evaluators. Under this allocation rule, each evaluator is responsible for 150 data points. Tthe evaluators are then asked to rate the quality of each data point on a scale from 1 (lowest) to 5 (highest). Our instructions to the evaluators are shown below:

Instructions to human evaluators.

We are curating a dataset partially generated by an AI model and want to seek feedback on its quality from human experts. During the evaluation process, we will provide both machine language sequences (the machine-generated operational sequences of experimental actions) and the corresponding natural language sequences (descriptions of experimental procedures in their original free texts).
You should rate these samples based on how well the operational sequences align with the original descriptions. Please use a rating scale of 1 (low alignment) to 5 (high alignment). Molecular skeletal formulas are provided as images for reference during evaluation. All original data for this dataset come from the United States Patent and Trademark Office (USPTO), ensuring the viability of the reactions.

The following are the detailed scoring guidelines, with a maximum score of 5:

- **5**: The machine-generated action sequence includes no errors in capturing key details of the original experimental procedure, including actions, materials, and numerical values.
- **4**: The machine-generated action sequence includes at most one ($n_{err} \leq 1$) error or omission related to actions, materials, or numerical values.
- **3**: The machine-generated action sequence includes at most two ($n_{err} \leq 2$) errors or omissions related to actions, materials, or numerical values.
- **2**: The machine-generated action sequence includes at most four ($n_{err} \leq 4$) errors or omissions related to actions, materials, or numerical values.
- **1**: The machine-generated action sequence includes more than four ($n_{err} > 4$) errors or omissions related to actions, materials, or numerical values.

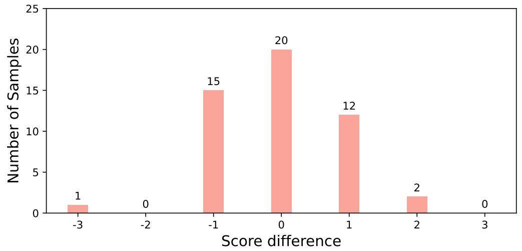


Figure 6: The score difference between evaluator 1 and evaluator 2 on 50 samples.

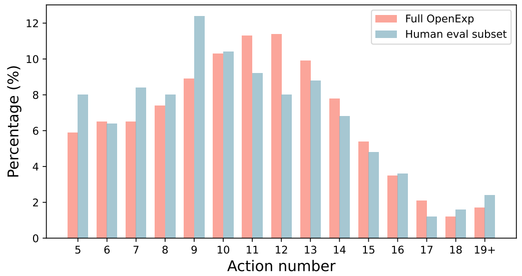


Figure 7: Action number distributions of the full OpenExp dataset and the human evaluation subset.

Figure 5 presents the human evaluation results. Statistics of these 250 data points and the entire dataset can be found in Figure 7 and Table 13. We can observe that the distribution of the sampled data points closely resembles that of the entire dataset, suggesting that the human evaluation results can reflect the overall quality of the OpenExp dataset.

Based on the 50 shared data points, we calculate the score differences in scores for the same samples (*i.e.,* , the score given by evaluator 1 minus the score given by evaluator 2). The results are presented in Figure 6. We can observe the exact alignment in ratings for 40% of the samples (20 out of 50), and a marginal score difference (±1) is recorded for 54% of the samples (27 out of 50). Discrepancies of two or more scores are exceedingly rare, occurring in just 6% of the samples (3 out of 50). Some examples of human evaluated data points are in Appendix C.3.2.

A.2 Collection and Preprocessing of ReactXT's Pretraining Dataset

In Section 3, we collect and compile a dataset to incrementally pretrain an LM for improved understanding of chemical reactions and individual molecules. Here we elaborate on the details of this dataset, which includes the following contents:

- A total of 1,162,551 chemical reactions;
- Patent abstracts and computed/experimental properties of 1,254,157 molecules, which are all

Property Name	Full OpenExp	Human eval.
Avg. Action Number	10.88	10.53
Avg. Reactant Number	2.96	2.98
Avg. Product Number	1.00	1.00
Avg. Catalyst Number	0.15	0.24
Avg. Solvent Number	1.06	1.00
Avg. Molecule Weight	164.88	168.80
Avg. Atom Number	10.57	10.69
Avg. Bond Number	10.63	10.75
Avg. Ring Number	1.06	1.07

Table 13: Chemical property statistics of the full OpenExp dataset and the human evaluation subset. Human eval stands for the human evaluation subset.

from the chemical reactions.

We extract chemical reactions from ORD and USPTO datasets. Then, we source patent abstracts from PubChem's Patent View[‖] and obtain molecular properties using the PubChem's PubView API[**]. For each molecule, the abstract text derives from the abstracts of patent documents where the molecule is mentioned, and its properties include both computational and experimental ones. Table 14 shows a complete list of these properties.

In Table 15, we compare the statistics of our pretraining dataset with that of PubChem324k. We can observe that ReactXT's pretraining dataset includes more molecules and additionally includes chemical reactions.

To prevent information leakage, we exclude a total of 54,403 reactions that appear in the validation and test sets of the downstream datasets (*i.e.,* OpenExp and USPTO-50K (Schneider et al., 2016)) from the pretraining dataset. The remaining 1,108,148 reactions are used for pretraining.

Discussion on License. The ORD database is accessible under the CC-BY-SA license, and the USPTO-Applications dataset is available under the CC0 license. The patent abstracts from PubChem are provided by Google Patent[††], which is released under the CC-BY-4.0 license. To comply with the strictest license terms, we will release our dataset under the CC-BY-SA license.

Additionally, we have utilized textual descriptions, computed properties, and experimental properties from the PubChem website for pretraining. Given that this data is aggregated from various sources by PubChem, determining a single appropriate license is challenging. To support future research while avoiding licensing complexities, we will provide the scripts for downloading and preprocessing this data, rather than distributing the data directly.

B Experimental Details

B.1 Hyperparameters

Here we detail the hyperparameters for ReactXT's pretraining and finetuning across three downstream tasks. Due to the prohibitive costs associated with training large LMs, finetuning on downstream datasets is limited to a single run.

ReactXT Pretrain. The pretraining stage of ReactXT has 5 million steps, with the number of molecules per reaction being $k = 4$. Following MolCA's (Liu et al., 2023b) experimental setup, we employ a Q-former with 8 query tokens. We use AdamW as the optimizer, with a weight decay set to 0.05. The optimizer's peak learning rate is set to 1×10^{-4}, scheduled by linear warmup with cosine decay. The warmup has 1000 steps and starts at a learning rate of 1×10^{-6}.

Experimental Procedure Prediction. We fully finetune all the baseline methods and ReactXT for 20 epochs, with a batch size of 32. The optimizer and learning rate settings are consistent with the pretraining phase.

Retrosynthesis. Following (Zhong et al., 2022), we sample 20 root-aligned augmentations for the training and testing subsets. Before finetuning on USPTO-50K, We first conduct 2 epochs of masked self-supervised pretraining for MolT5 and ReactXT on the USPTO-full dataset (Dai et al., 2019), following the pretraining strategy of R-SMILES (Zhong et al., 2022). During finetuning, we train MolT5 for 20 epochs and ReactXT for 5 epochs on the augmented training set using a batch size of 32. We then average the model's parameters on the last several tuning steps as the final checkpoint for testing. During testing, we conduct a beam search with a beam size of 20 for both models and return the top ten results as the model's predictions. The beam size (20) and the number of results (10) are following the experiment of R-SMILES (Zhong et al., 2022). The optimizer and learning rate settings are kept consistent with the pretraining phase.

Molecule Captioning. On both datasets, we full finetune MolCA and ReactXT 20 epochs, with a batch size of 32. The optimizer and learning rate settings are consistent with the pretraining phase.

[‖] pubchem.ncbi.nlm.nih.gov/docs/patents

[**] pubchem.ncbi.nlm.nih.gov/docs/pug-view

[††] patents.google.com

Computed Properties		Experimental Properties					
Property	Count	Property	Count	Property	Count	Property	Count
Molecular Weight	1244109	Physical Description	8368	Vapor Density	1043	Enthalpy of Sublimation	9
Hydrogen Bond Donor Count	1244109	Kovats Retention Index	6878	Autoignition Temperature	771	Acid Value	4
Hydrogen Bond Acceptor Count	1244109	Solubility	5909	Heat of Vaporization	583	Dielectric Constant	2
Rotatable Bond Count	1244109	Chemical Classes	5726	Viscosity	550	Dispersion	1
Exact Mass	1244109	Melting Point	4468	Taste	514	Hydrophobicity	1
Monoisotopic Mass	1244109	Vapor Pressure	3032	Henry's Law Constant	502		
Topological Polar Surface Area	1244109	Boiling Point	2996	Surface Tension	448		
Heavy Atom Count	1244109	Color/Form	2927	pH	444		
Formal Charge	1244109	Density	2862	Odor Threshold	442		
Complexity	1244109	LogP	2763	Corrosivity	410		
Isotope Atom Count	1244109	Other Experimental Properties	2393	Heat of Combustion	405		
Defined Atom Stereocenter Count	1244109	Decomposition	2033	Ionization Efficiency	332		
Undefined Atom Stereocenter Count	1244109	Refractive Index	1777	Optical Rotation	265		
Defined Bond Stereocenter Count	1244109	Collision Cross Section	1634	Ionization Potential	253		
Undefined Bond Stereocenter Count	1244109	Odor	1512	LogS	166		
Covalently-Bonded Unit Count	1244109	Stability/Shelf Life	1506	Polymerization	134		
Compound Is Canonicalized	1244109	Flash Point	1479	Relative Evaporation Rate	101		
XLogP3	1184175	Dissociation Constants	1250	Caco2 Permeability	79		

Table 14: Statistics of the collected molecule properties, including computed properties and experimental properties.

	Our Dataset	Pubchem324k
Num of Molecules	1,254,157	313,083
Num of Reactions	1,162,551	-
Avg. Molecule Weight	362.4	502.4
Avg. Atom Count	24.9	35.2
Avg. Bond Count	26.8	37.6
Avg. Ring Count	2.9	3.5
Avg. Text Length	517.8	120.4
Avg. Property Count	17.8	-

Table 15: Statistics of ReactXT's pretraining dataset and Pubchem324k.

B.2 Other Implementation Details

Baselines. We briefly introduce the baselines:

- **Galactica** (Taylor et al., 2022). Galactica is a scientific language model which is pretrained on 2 million compounds from PubChem. It has a decent understanding of SMILES formulas.

- **MolT5** (Edwards et al., 2022). MolT5 is developed based on the T5 model. Its training corpora include both natural language and SMILES data, making it suitable for both molecule captioning and text-based molecular generation tasks.

- **TextChemT5** (Christofidellis et al., 2023). TextChemT5 is a T5-based multi-domain LM, which is tuned on various text-molecule tasks.

- **MolCA** (Liu et al., 2023b). MolCA is a multimodal language model finetuned on Galactica. It includes both graph encoder and LM, where a Querying Transformer is applied to align their latent spaces.

- **AT** (Tetko et al., 2020). AT trains transformers with data augmentation for retrosynthesis. The data augmentation is achieved by rearranging the order of characters in SMILES strings in both the training and test sets.

- **MEGAN** (Sacha et al., 2021). MEGAN represents chemical reactions as a sequence of graph edits and performs retrosynthesis by sequentially modifying the target molecule.

- **MoMu** (Su et al., 2022). Momu contrastively pretrains a GNN and an LM with paired molecular graph-text data, and can be adapted to retrieval and generation tasks.

Pretrain Input Context	Pretrain Data Type	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
No incremental pretrain	-	39.4	32.2	52.7	39.4	47.6	49.2
Reactions	reaction	37.3	29.9	50.3	36.5	45.0	46.7
ReactXT	reaction, sing. mol.	**42.6**	**35.2**	**54.7**	**41.7**	**49.6**	**51.2**

Table 16: Ablation study. Performances (%) for molecule captioning on the PubChem324k dataset.

	BLEU2	BLEU4	ROUGE1	ROUGE2	ROUGEL
T-statistic	14.619	13.622	16.126	14.438	15.053
P-value	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 17: P-values for experimental procedure prediction (Table 5), comparing ReactXT against MolCA-1.3B.

	BLEU2	BLEU4	ROUGE1	ROUGE2	ROUGEL	METEOR
T-statistic	3.469	3.823	3.451	3.851	3.434	4.107
P-value	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**

Table 18: P-values for captioning on PubChem324k (Table 6), comparing ReactXT against MolCA-1.3B, full ft.

	BLEU2	BLEU4	ROUGE1	ROUGE2	ROUGEL	METEOR
T-statistic	2.918	3.523	2.843	3.495	3.129	2.195
P-value	**0.004**	**<0.001**	**0.004**	**<0.001**	**0.002**	**0.028**

Table 19: P-values for captioning on CheBI-20 (Table 6), comparing ReactXT against MolCA-1.3B, full ft.

	Top1	Top3	Top5	Top10
Z-statistic	2.340	2.380	0.440	0.000
P-value	**0.019**	**0.017**	0.662	1.000

Table 20: P-values for retrosynthesis (Table 7), comparing ReactXT against R-SMILES. Both models use 20 augmentations during testing.

- **Chemformer** (Irwin et al., 2022). Chemformer is a Transformer-based molecule LM that is self-supervised pretrained on a SMILES corpus. It can be applied to both generation and property prediction tasks.
- **Random, among all reactions** (Vaucher et al., 2021). Randomly pick an action sequence from the training set.
- **Random, compatible pattern** (Vaucher et al., 2021). Randomly pick an action sequence from the training subset of reactions that have the same number of molecules as the current reaction.
- **Nearest Neighbor** (Vaucher et al., 2021). Pick the action sequence from the training set with the reaction most similar to the current one, as determined by reaction fingerprints (Schwaller et al., 2019).

C More Experimental Results

C.1 Ablation Study

Table 16 presents an ablation study examining the impact of input contexts on molecule captioning. The removal of the random molecule context results in diminished captioning performance. This observation can be attributed to two factors: 1) including the PubChem324k dataset, which is used for creating random molecule contexts, is important to maintain molecule captioning performance; and 2) without random molecule contexts, the LM becomes overly dependent on reaction contexts, compromising its capability to accurately caption individual molecules. This finding underscores the significance of incorporating random molecule contexts in training.

C.2 Statistical Analysis

We carry out statistical tests on the experimental results to demonstrate that ReactXT achieves a significant performance improvement compared to the baseline models. For most metrics (such as BLEU, ROUGE, METEOR), we employ the T-test; for Top-k accuracy, where calculating the standard deviation was challenging, we use a 2-proportion Z-test instead.

The results of the statistical tests are presented in Tables 17 to 20. We **bold** p-values that are smaller than 0.05. From these tables, it can be observed that our method achieves statistically significant improvements across all metrics within the tasks of experimental procedure prediction (Table 17) and molecule captioning (Tables 18 and 19). As for the retrosynthesis task (Table 20), our method demonstrates statistically significant enhancements in both Top1 and Top3 accuracies. These observations collectively demonstrate the effectiveness of our proposed pretraining method.

Field	Value		
Reactant	1: OCCCCCCCc1ccccc1 2: C#CC(=O)O 4: c1ccccc1		
Catalyst	3: Cc1ccc(S(=O)(=O)O)cc1		
Product	-1: C#CC(=O)OCCCCCCCc1ccccc1		
Source	A mixture of 0.5 g of 7-phenylheptanol, 0.27 g of propiolic acid, 0.005 g of p-toluenesulfonic acid and 25 ml of benzene was refluxed with stirring for six hours while water formed was removed by a Dean-Stark water separator. After the reaction was completed, the reaction solution was washed successively with a 5% aqueous sodium bicarbonate solution and a saturated sodium chloride solution, and dried over anhydrous magnesium sulfate. After removal of the solvent under reduced pressure, the obtained residue was subjected to silicagel column chromatography to yield 0.368 g of 7-phenylheptyl propiolate (compound 3).		
Annotated Actions	MAKESOLUTION with 1 (1.1 g) and 2 (0.005 g) and 3 (25 ml) and 4 ; REFLUX for 6 hours ; CONCENTRATE ; WASH with NaHCO3 ; WASH with sodium chloride ; DRYSOLUTION over magnesium sulfate ; FILTER keep filtrate ; YIELD -1 (1.15 g).	Predicted Actions	MAKESOLUTION with 1 (0.27 g) and 2 (0.005 g) and 3 (25 ml) and 4 ; REFLUX for 10 hours; CONCENTRATE ; WASH with NaHCO3 ; WASH with sodium chloride ; DRYSOLUTION over magnesium sulfate ; FILTER keep filtrate ; YIELD -1 (0.368 g).

(a) Example 1.

Field	Value		
Reactant	1: C[Si]1(C)CC[Si](C)(C)N1c1ccc(C(O)c2cn(S(=O)(=O)c3ccccc3)c3ncc(Cl)cc23)cn 2: Nc1ccc(C(O)c2cn(S(=O)(=O)c3ccccc3)c3ncc(Cl)cc23)cn1 4: CC[SiH](CC)CC 5: O=C(O)C(F)(F)F		
Solvent	3: ClCCl		
Product	-1: Nc1ccc(Cc2cn(S(=O)(=O)c3ccccc3)c3ncc(Cl)cc23)cn1		
Source	To (1-benzenesulfonyl-5-chloro-1H-pyrrolo[2,3-b]pyridin-3-yl)-[6-(2,2,5,5-tetramethyl-[1,2,5]azadisilolidin-1-yl)-pyridin-3-yl]-methanol and (6-amino-pyridin-3-yl)-(1-benzenesulfonyl-5-chloro-1H-pyrrolo[2,3-b]pyridin-3-yl)-methanol (118, 119, 1.70/1.25 g mix, 2.41 mmol) in 25.0 mL of dichloromethane, triethylsilane (3.00 mL, 18.8 mmol) and trifluoroacetic acid (1.50 mL, 19.5 mmol) were added and the reaction stirred at room temperature overnight. The reaction was concentrated under vacuum, combined with aqueous potassium carbonate and extracted with ethyl acetate. The organic layer was dried over sodium sulfate, filtered and the filtrate concentrated under vacuum. The resulting material was purified by silica gel column chromatography eluting with 20-100% ethyl acetate in hexane to provide the desired compound (120, 0.70 g).		
Annotated Actions	MAKESOLUTION with 1 and 2 and 3 (25.0 mL) ; ADD 4 (3.00 mL, 18.8 mmol) ; ADD 5 (1.50 mL, 19.5 mmol) ; STIR for overnight at room temperature ; CONCENTRATE ; ADD K2CO3 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1 (0.70 g).	Predicted Actions	MAKESOLUTION with 1 (1.00 g, 1.91 mmol) and 2 (0.69 g, 1.72 mmol) and 3 (35 mL) ; ADD 4 (1.35 mL, 7.84 mmol) ; ADD 5 (1.90 mL, 26.7 mmol) ; STIR for 8 h at room temperature ; CONCENTRATE ; EXTRACT with K2CO3 ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1 (0.13 g, 19%).

(b) Example 2.

Table 21: Examples of accurate experimental procedure predictions.

C.3 Case Studies and Error Analysis

C.3.1 Experimental Procedure Prediction

In this section, we present case studies from the experimental procedure prediction task to inform future research. We include examples of accurate predictions (see Table 21), inaccurate predictions (see Tables 22), and predictions that are different from the annotations but may also work (see Table 23 and Table 24). Our selection criteria prioritizes the accuracy of action sequences and the correct identification of primary materials, while overlooking specifics like material quantities and temperatures. All the examples are from the test set of OpenExp.

Table 21 displays two examples where experimental procedures are accurately predicted, showing close alignment between predicted and annotated actions, albeit with slight variances in material quantities and experiment times. These cases highlight the capability of LMs to predict experimental procedures, suggesting a path toward automating chemical synthesis.

Table 22 displays two failed examples of experimental procedure prediction. The predicted action sequences significantly deviate from the annotated sequences, making them impractical. Additionally, we can observe one common error of repetition, with the same or similar actions being duplicated.

Tables 23 and Table 24 showcase three examples where the predictions, while different from the annotations, could still be viable. In Example 5, as an alternative to the annotated 'EXTRACT with ethyl acetate', the model proposes a series of actions ('COLLECT LAYER', 'WASH with ethyl acetate', 'DRY SOLUTION', and 'FILTER'), serving a similar function. In Example 6, instead of the specified 'SET TEMPERATURE' and 'STIR', the model recommends 'STIR for 1h at 0 °C', serving the same purpose. In Example 7, the model suggests adding components ('ADD 4', 'ADD 5', 'ADD 6') sequentially rather than making a single solution as annotated, which could also be effective.

C.3.2 Human Evaluation of OpenExp

In this section, we present case studies from human evaluations on the OpenExp dataset. Samples rated from 5 to 1 by human evaluators are included, as shown in Tables 25 to Tables 29. All samples are from the 250 human evaluated data points (see Appendix A.1). It can be observed that samples with two or fewer errors may only have minor flaws, such as typol errors or incorrect numerical values.

Field	Value		
Reactant	1: Nc1ccc(C(=O)N[C@H](CO)Cc2ccccc2)c(/C=C/c2ccccc2)c1 3: CC(=O)OC(C)=O		
Solvent	2: C1CCOC1		
Product	-1: CC(=O)Nc1ccc(C(=O)N[C@H](CO)Cc2ccccc2)c(/C=C/c2ccccc2)c1		
Source	1 g (2.7 mmol) of (S)-4-amino-2(E-2-phenylethen-1-yl)-N-(3-phenylpropan-1-ol-2-yl)benzamide (intermediate 43f) was suspended in 50 ml of tetrahydrofuran and mixed with 0.25 ml (2.7 mmol) of acetic anhydride at 100° C. The mixture was stirred for 16 h. The reaction was then concentrated under reduced pressure and the residue was recrystallized from ethanol. 0.78 g (71%) of the product was obtained.		
Annotated Actions	ADD 1 (1 g, 2.7 mmol) ; ADD 2 (50 ml) ; ADD 3 (0.25 ml, 2.7 mmol) ; STIR for 16 h at 100 °C ; CONCENTRATE ; RECRYSTALLIZE from ethanol (0.78 g) ; YIELD -1.	Predicted Actions	MAKESOLUTION with 1 and 2 (15 ml) ; ADD 3 (0.27 ml, 2.9 mmol) ; STIR for 10 h at 100 °C ; WAIT for 10 h at 100 °C ; STIR for 10 h at 100 °C ; CONCENTRATE ; ADD ethyl acetate ; ADD water ; PHASESEPARATION ; COLLECTLAYER aqueous ; EXTRACT with ethyl acetate ; COLLECTLAYER organic ; WASH with brine ; DRYSOLUTION over magnesium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1 (144 mg, 75%).

(a) Example 3.

Field	Value		
Reactant	1: Brc1ccc2noc(-c3ccccc3)c2c1 2: O		
CATALYST	3: [Zn]		
SOLVENT	4: CC(=O)O		
PRODUCT	-1: Nc1ccc(Br)cc1C(=O)c1ccccc1		
Source	5-Bromo-3-phenyl-2,1-benzisoxazole (7.5 g, 28.6 m mol), water (14.6 ml), and zinc dust (9.3 g, 143 m mol) were combined. Acetic acid (8.6 ml, 143 m mol) was added and the mixture was stirred and heated at 80° C. for 90 minutes. After cooling to room temperature, both the liquid and solid portion of the reaction were extracted with methylene chloride. The combined methylene chloride solutions was washed once with sodium hydroxide solution (10%) and several times with water. Drying (sodium sulfate) and evaporation yielded the desired product (7.42 g) of mp. 92°-102° C.		
Annotated Actions	ADD 1 (7.5 g) ; ADD 2 (14.6 ml) ; ADD 3 (9.3 g) ; ADD 4 (8.6 ml) ; STIR for 90 minutes at 80 °C ; EXTRACT with methylene chloride ; EXTRACT with CH2 Cl2 ; WASH with NaOH ; WASH with 2 ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; YIELD -1 (7.42 g).	Predicted Actions	MAKESOLUTION with 1 and 2 (150 ml) ; ADD 3 (24 g) ; ADD 4 (60 ml) ; STIR for 120 minutes at 80 °C ; WAIT for 120 minutes at 80 °C ; FILTER ; ADD carbon ; COLLECTLAYER organic ; WASH with 2 (200 ml) 3 x ; DRYSOLUTION over sodium sulfate ; FILTER keep filtrate ; CONCENTRATE ; RECRYSTALLIZE from 2-amino-5-bromo-benzophenone (20.7 g) ; YIELD -1 (20.7 g, 57.9%).

(b) Example 4.

Table 22: Examples of inaccurate experimental procedure predictions. Green denotes error of repetition.

Field	**Value**		
Reactant	1: COc1ccc(-c2cccc(CC(=O)O)c2)cc1C(C)=NO 3: FC(F)(F)c1ccc(CBr)cc1 4: [H-].[Na+] 5: Cl		
SOLVENT	2: C1CCOC1		
PRODUCT	-1: COc1ccc(-c2cccc(CC(=O)O)c2)cc1C(C)=NOCc1ccc(C(F)(F)F)cc1		
Source	9 mg of [3'-(1-Hydroxyiminoethyl)-4'-methoxybiphenyl-3-yl]acetic acid was dissolved in 0.2 mg of tetrahydrofuran, then 25 mg of 4-trifluoromethylbenzyl bromide and 10 mg of sodium hydride were added thereto, and the mixture was stirred at room temperature for 3 days. The reaction mixture was neutralized with 1N hydrochloric acid and then extracted with ethyl acetate. The organic layer was concentrated, and the residue was purified by reverse-phase high performance liquid chromatography to give 5.21 mg of the title compound.		
Annotation Actions	ADD 1 (9 mg) ; ADD 2 (0.2 mg) ; ADD 3 (25 mg) ; ADD 4 (10 mg) ; STIR for 3 days at room temperature ; PH with 5 (1 normal) to pH neutral at room temperature ; EXTRACT with ethyl acetate ; CONCENTRATE ; YIELD -1 (5.21 mg).	Predicted Actions	MAKESOLUTION with 1 (25 mL) and 2 ; ADD 3 (292 mg) ; ADD 4 (104 mg) ; STIR for overnight at room temperature ; PH with 5 (1 normal) to pH neutral at room temperature ; COLLECTLAYER organic ; WASH with ethyl acetate ; COLLECTLAYER organic ; WASH with water ; WASH with brine ; DRYSOLUTION over magnesium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD -1 (204 mg).

(a) Example 5.

Field	**Value**		
REACTANT	3: Cc1ccc2c(N)ccc(O)c2n1 4: O=N[O-].[Na+] 5: [N-]=[N+]=[N-].[Na+]		
Solvent	1: Cl 2: O		
PRODUCT	-1: Cc1ccc2c(N=[N+]=[N-])ccc(O)c2n1		
Source	5-Amino-8-hydroxy-2-methylquinoline (12; 723 mg, 4.2 mmol) was dissolved in a solution of concentrated hydrochloric acid (0.4 mL) and water (5 mL), cooled to -3° C. in a salt-ice bath, stirred for 10 min, then treated dropwise with a cold solution of sodium nitrite (0.50 g, 7.2 mmol) in water (5 mL). The mixture was stirred for 20 min, then treated dropwise with sodium azide (0.60 g, 9.2 mmol) in water (40 mL), stirred at 0° C. for a further 1.5 h, then allowed to warm to room temperature over 24 h in the dark. Isolation by extraction with diethyl ether gave a dark brown solid, which was recrystallized from light petroleum to yield 13 as light brown crystals (554 mg, 66%). 1H NMR (500 MHz, CD3OD) δ ppm: 3.31 (s, 3H), 7.06 (d, J=8.5 Hz, 1H), 7.15 (d, J=8.5 Hz, 1H), 7.36 (d, J=8.5 Hz, 1H), 8.21 (d, J=8.5 Hz, 1H). 13C NMR (125 MHz, CD3OD) δ ppm: 24.9, 111.8, 115.6, 121.6, 123.7, 127.9, 132.6, 139.6, 151.1, 159.7. HRMS (ESI): calcd for C10H8N4O [M+H]+: 201.0771, found: 201.0773.		
Annotation Actions	MAKESOLUTION with 1 (0.4 mL) and 2 (5 mL) ; ADD 3 (723 mg, 4.2 mmol) ; SETTEMPERATURE -3 °C ; STIR for 10 min ; MAKESOLUTION with 4 (0.50 g, 7.2 mmol) and 2 (5 mL) ; STIR for 20 min ; MAKESOLUTION with 5 (0.60 g, 9.2 mmol) and 2 (40 mL) ; STIR for 1.5 h at 0 °C ; RECRYSTALLIZE from petroleum ; YIELD -1 (554 mg, 66%).	Predicted Actions	MAKESOLUTION with 1 (3 ml) and 2 (1 ml) ; ADD 3 (210 mg, 1.24 mmol) ; STIR for 1 h at 0 °C; MAKESOLUTION with 4 (80 mg, 1.24 mmol) and 2 (1 ml) ; STIR for 1 h at 0 °C; MAKESOLUTION with 5 (88 mg, 1.24 mmol) and 2 (1 ml) ; STIR for 2 h at 0 °C ; RECRYSTALLIZE ; YIELD -1 (120 mg, 47%).

(b) Example 6.

Table 23: Examples of experimental procedure predictions that are different from the annotation but might be viable.

<table>
<tr><th>Field</th><th colspan="3">Value</th></tr>
<tr><td>Reactant</td><td colspan="3">$1$: CNC(=O)c1cn(CCCCc2ccc(N)nn2)nn1
$2$: O=C(O)Cc1cc(Br)ccn1
$4$: CCCP1(=O)OP(=O)(CCC)OP(=O)(CCC)O1
$6$: CCN(C(C)C)C(C)C</td></tr>
<tr><td>Solvent</td><td colspan="3">$3$: CN(C)C=O
$5$: CCOC(C)=O
$7$: O</td></tr>
<tr><td>Product</td><td colspan="3">$-1$: CNC(=O)c1cn(CCCCc2ccc(NC(=O)Cc3cc(Br)ccn3)nn2)nn1</td></tr>
<tr><td>Source</td><td colspan="3">To a solution of 1-(4-(6-aminopyridazin-3-yl)butyl)-N-methyl-1H-1,2,3-triazole-4-carboxamide (2.0 g, 7.2 mmol) and 2-(4-bromopyridin-2-yl)acetic acid (1.6 g, 7.2 mmol) in DMF (5 ml) were added T3P (50 wt. % in EtOAc, 9.2 g, 14 mmol) and DIEA (2.7 g, 22 mmol). The mixture was stirred at RT for 3 h, then water (50 mL) was added, the mixture was stirred for 30 min, and precipitate was isolated by filtration to give the title compound as a yellow solid (1 g, 40%). MS (ES+) C19H21BrN8O2 requires: 473. found: 474 [M+H]+.</td></tr>
<tr><td>Annotated Actions</td><td>MAKESOLUTION with $1$ (2.0 g, 7.2 mmol) and $2$ (1.6 g, 7.2 mmol) and $3$ (5 ml) ;
MAKESOLUTION with $4$ and $5$ (9.2 g, 14 mmol) and $6$ (2.7 g, 22 mmol) ;
STIR for 3 h at RT ;
ADD $7$ (50 mL) ;
STIR for 30 min ;
YIELD $-1$ (1 g, 40%).</td><td>Predicted Actions</td><td>MAKESOLUTION with $1$ (30 mg, 0.109 mmol) and $2$ (23 mg, 0.109 mmol) and $3$ (0.5 ml) ;
ADD $4$ (67 l, 0.109 mmol, 0.109 mmol) ;
STIR for 3 h at room temperature ;
ADD $5$ (5 ml) ;
ADD $6$ (67 l, 0.38 mmol, 0.38 mmol) ;
STIR for 30 min at room temperature ;
ADD $7$ (0.5 ml) ;
CONCENTRATE ;
YIELD $-1$ (36 mg, 70% yield).</td></tr>
</table>

Table 24: Examples of experimental procedure predictions that are different from the annotation but might be viable. Example 7.

Field	Value
Source	To a solution of 2-(5-amino-3-methyl-1H-pyrazol-4-yl)-benzothiazole-5-carboxylic acid ethyl ester (30 mg) in THF (1 mL) was added lithium aluminum hydride (4 mg). The reaction mixture was stirred at room temperature for 5 hrs at which point sodium sulfate nonahydrate was added. The resulting mixture was stirred for an additional 30 min. The solids were removed by filtration. The solvent was then evaporated and the residue was purified by flash column chromatography eluting with CHCl3:MeOH=9:1 to yield 21 mg (81%) of the title compound as a cream coloured solid. MS (m/z, ES+): 261.1 (M+1, 100%).
Annotated Actions	MAKESOLUTION with 2-(5-amino-3-methyl-1H-pyrazol-4-yl)-benzothiazole-5-carboxylic acid ethyl ester (30 ; mg) and THF (1 mL) ; ADD lithium aluminum hydride (4 mg) ; STIR for 5 hr at room temperature ; ADD sodium sulfate nonahydrate ; STIR for 30 min at room temperature ; FILTER keep filtrate ; CONCENTRATE ; YIELD PRODUCT (21 mg, 81%) .

Table 25: Example with a Human Evaluation Score of 5. The action sequence accurately captures the source paragraph.

Field	Value
Source	A mixture of (5-nitro-pyridin-2-yl)-(2,2,2-trifluoro-ethyl)-amine (230 mg, 1.04 mmol), cesium carbonate (730 mg, 2.07 mmol) and iodomethane (0.59 mL, 4.18 mmol) in DMF (4 mL) was heated in a sealed tube at 50° C. for 3 hr. The reaction mixture was evaporated to dryness and the crude was partitioned between methylene chloride and water. The organic layer was dried over magnesium sulfate, filtered and concentrated to give methyl-(5-nitro-pyridin-2-yl)-(2,2,2-trifluoro-ethyl)-amine (270 mg, crude) as a brown solid, which was directly used in the next step reaction without further purification. LCMS calcd for C8H8F3N3O2 (m/e) 235, obsd 236 (M+H).
Annotated Actions	MAKESOLUTION with (5-nitro-pyridin-2-yl)-(2,2,2-trifluoro-ethyl)-amine (230 mg, 1.04 mmol) and cesium carbonate (730 mg, 2.07 mmol) and iodomethane (0.59 mL, 4.18 mmol) and DMF (4 mL) ; STIR for 3 hr at 50 °C ; CONCENTRATE ; PARTITION with methylene chloride and water ; COLLECTLAYER organic ; DRYSOLUTION over magnesium sulfate ; FILTER keep filtrate ; CONCENTRATE ; YIELD PRODUCT (270 mg) .

Table 26: Example with a Human Evaluation Score of 4. The action sequence contains 1 error, which is highlighted in green.

Field	Value
Source	To a stirred solution of 1,5-anhydro-2,3-dideoxy-D-erythro-hexitol (44.9 g) and imidazole (65.2 g) in DMF (500 ml) was added tert-butylchlorodiphenylsilane (88.5 mL) at 0° C. After stirring for 4 h, the reaction mixture was diluted with EtOAc (1000 ml). The organic layer was washed with water (200 mL×5) and brine (200 mL), and dried over Na2SO4. The solution was concentrated under reduced pressure, and the residue was purified by column chromatography (PE/EtOAc) to afford 70.9 g of the title compound as a colorless oil.
Annotated Actions	MAKESOLUTION with 1,5-anhydro-2,3-dideoxy-D-erythro-hexitol (44.9 g) and imidazole (65.2 g) and DMF (500 ml) ; ADD tert-butylchlorodiphenylsilane (88.5 mL) at 0 °C ; ADD tert-butylchlorodiphenylsilane (88.5 mL) ; STIR for 4 h at 0 °C ; ADD ethyl acetate (1000 ml) ; COLLECTLAYER organic ; WASH with water (200 mL) ; WASH with brine (200 mL) ; DRYSOLUTION over Na2SO4 ; FILTER keep filtrate ; CONCENTRATE ; YIELD PRODUCT (70.9 g) .

Table 27: Example with a Human Evaluation Score of 3. The action sequence contains 2 errors, which are highlighted in green.

Field	**Value**
Source	A mixture of methyl 3-hydroxy-1-methyl-1H-pyrazole-5-carboxylate (2.34 g, 15.0 mmol), iodomethane (3.19 g, 22.5 mmol), potassium carbonate (4.15 g, 30.0 mmol) and N,N-dimethylformamide (15 ml) was stirred at room temperature for 18 hr. The mixture was diluted with water (50 mL), and extracted with ethyl acetate (50 mL×3). The organic layer was washed with water (10 mL×2), and concentrated under reduced pressure. The residue was purified by silica gel column chromatography (hexane/ethyl acetate=100/0→50/50) to give the title compound (2.01 g, yield 79%) as a white solid. 1H-NMR (DMSO-d6, 300 MHz) 3.78 (3H, s), 3.81 (3H, s), 3.94 (3H, s), 6.27 (1H, s).
Annotated Actions	MAKESOLUTION with methyl 3-hydroxy-1-methyl-1H-pyrazole-5-carboxylate (2.34 g, 15.0 mmol) and iodomethane (3.19 g, 22.5 mmol) and potassium carbonate (4.15 g, 30.0 mmol) and N,N-dimethylformamide (15 ml) ; STIR for 18 hr at room temperature ; ADD water (50 mL) at room temperature over 18 hr ; COLLECTLAYER organic ; WASH with ethyl acetate (50 mL) ; COLLECTLAYER organic ; WASH with water (10 mL) ; CONCENTRATE ; YIELD PRODUCT (2.01 g, yield 79%) .

Table 28: Example with a Human Evaluation Score of 2. The action sequence contains 4 errors, which are highlighted in green.

Field	**Value**
Source	4-[4-(4-Fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid Sodium hydroxide (40 mg, 1 mmol) was added to a suspension of 4-[4-(4-fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid methyl ester (40 mg) in a mixture of water (1 mL) and dioxane (1 mL). The resulting mixture was heated at 50° C. for 4 h. The solvent was evaporated and water (5 mL) was added. The mixture was filtered and the filtrate was made acidic to pH 3 by the addition of concentrated HCl. The precipitate was collected by filtration and dried to give 4-[4-(4-fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid (45 mg, 58% for two steps). 1H NMR (300 MHz, DMSO-d6) 13.21 (br s, 1H), 8.65 (s, 1H), 8.17-8.33 (m, 5H), 7.86 (t, J=7.6 Hz, 1H), 7.73 (t, J=7.8 Hz, 1H), 7.52 (t, J=8.5 Hz, 2H), 7.40 (t, J=8.6 Hz, 2H).
Annotated Actions	MAKESOLUTION with 4-[4-(4-fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid methyl ester (40 mg) and 4-[4-(4-fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid methyl ester (40 mg) and water (1 mL) and dioxane (1 mL) ; ADD 4-[4-(4-Fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid Sodium hydroxide (40 mg, 1 mmol) ; MAKESOLUTION with water (1 mL) and dioxane (1 mL) ; STIR for 4 h at 50 °C ; CONCENTRATE ; ADD water (5 mL) ; FILTER keep precipitate ; {missing operation} ; DRYSOLUTION over 4-[4-(4-fluoro-phenyl)-thiazol-2-yl]-2'-nitro-biphenyl-2-carboxylic acid (45 mg) ; FILTER keep filtrate ; YIELD PRODUCT (45 mg, 58%) .

Table 29: Example with a Human Evaluation Score of 1. The action sequence contains 5 errors, which are highlighted in green.