BATGPT-CHEM: A FOUNDATION LARGE MODEL FOR CHEMICAL ENGINEERING

Yifei Yang[♠], **Runhan Shi**[♠], **Zuchao Li**[♣], **Shu Jiang**[♢], **Yang Yang**[♠], **Bao-Liang Lu**[♠], **Hai Zhao**[♠]*
[♠]Shanghai Jiao Tong University [♣]Wuhan University [♢]Nantong University
{yifeiyang, han.run.jiangming}@sjtu.edu.cn zcli-charlie@whu.edu.cn jshmjs45@gmail.com
bllu@sjtu.edu.cn {yangyang, zhaohai}@cs.sjtu.edu.cn

ABSTRACT

LLMs have showcased remarkable capabilities in the realm of AI for Science (Ai4Sci) and the chemistry has greatly benefited from the advancement of AI tools. With a strong capacity for learning sequential data like natural language, LLMs offer immense potential. Notably, common representations in chemistry, such as SMILES, are also in the form of sequences. Hence, we propose leveraging LLMs to comprehensively model both chemical sequences and natural language sequences, aiming to tackle diverse chemical tasks. To fulfill this objective, we introduce BatGPT-Chem, a foundational large-scale model with 15B parameters tailored for chemical engineering. First, we unify diverse tasks in chemistry by modeling them through a combination of natural language and SMILES. Next, leveraging this unified modeling approach, we craft prompt templates and generate instructional tuning data using a substantial volume of chemical data. Subsequently, we train BatGPT-15B on over a hundred million instances of instructional tuning data, empowering it to address tasks such as **Molecule Description**, **Molecule Design**, **Retro-synthesis Prediction**, **Product Inference**, and **Yield Prediction**. We release our trial platform at `https://www.batgpt.net/dapp/chem`.

1 INTRODUCTION

The continuous advancement of artificial intelligence (AI), particularly large language models (LLMs), has empowered machines to excel in various domains. Natural language serves as the most intuitive means of communication and expression, rendering the path for AI applications remarkably smooth as long as effective natural language understanding is achieved. The high intelligence of LLMs is regarded as a key element in overcoming the "AI4Science" challenge in recent years.

The rapid advancement of the field of chemistry is inseparable from the progress in computer-assisted synthesis technology and automated management of chemical knowledge. In the past few decades, various computer-assisted synthesis algorithms or softwares based on reaction templates have been proposed (Corey & Wipke, 1969; Salatin & Jorgensen, 1980; Smith & Sherwood, 1976; Dugundji & Ugi, 2006). However, these reaction templates are expert-developed manual rules, which not only incur significant manpower and time costs but also fail to cover all complex organic chemistry prediction problems. Databases such as Reaxys [1], SciFinder [2], ChemSpider [3], and SPRESI [4] have aided chemists in searching for literature sources or similar reaction instances. However, the full potential of modern computers has still not been realized, as reaction space searches still require manual intervention by chemists. Subsequently, a multitude of neural network-driven chemical prediction algorithms (Rappoport et al., 2014; Wei et al., 2016; Simm & Reiher, 2017; Xie et al., 2021; Edwards et al., 2021) have emerged, contributing to the advancement of AI in the field of chemistry. However, these methods often specialize in completing specific chemical tasks, such as singular product prediction or molecular description.

* Corresponding author.

[1] `https://www.reaxys.com/`
[2] `https://scifinder-n.cas.org/`
[3] `http://www.chemspider.com/`
[4] `https://www.spresi.com/`

Recently, although a small number of studies have applied state-of-the-art LLMs to the chemistry (Boiko et al., 2023; Qian et al., 2023; Li et al., 2023b), they have not adequately retrained LLMs using chemical data. Since chemical symbols are often regarded as a specialized language, models trained on large-scale natural language datasets struggle to fully comprehend various chemical symbols.

In this work, we consider the widely used SMILES notation in chemistry as a specialized language and employ unified modeling to integrate it with natural language using LLMs. We design multiple instruction tuning tasks and convert various open-source and our closed-source datasets into a large-scale instruction tuning dataset using prompt templates. Building upon our team's BatGPT-15B (Li et al., 2023c), we expand its vocabulary with additional chemical terms and instruction tune it, culminating in the model of BatGPT-Chem: A Foundational Large Model for Chemical Engineering. This LLM surpasses existing scarce large chemical models (Zhang et al., 2024; Zhao et al., 2024) in model size and utilizes a larger-scale instruction fine-tuning dataset in both Chinese and English. Additionally, we utilize more data to enhance the model's molecular design capabilities.

2 Related Works

The advancement of chemistry in recent decades has been closely intertwined with the support of computer and AI technologies. The development of AI applications in the field of chemistry can be broadly categorized into rule-based chemistry AI systems, neural network and small language model-based chemistry AI systems, and LLM-based chemistry AI systems.

Rule-based chemistry AI systems. Over the past few decades, there have been numerous rule-based and template-based chemistry AI systems (Corey & Wipke, 1969; Salatin & Jorgensen, 1980; Smith & Sherwood, 1976; Davis & King, 1984; Dugundji & Ugi, 2006; Lu et al., 2005). During development, these systems entail significant manual design of reaction templates, involving expert computational chemists, resulting in a challenging design process. They also require users to input compound details and reaction conditions manually, which incurs a high learning cost. Additionally, numerous chemical databases have also been proposed, such as Reaxys, CDS (Fletcher et al., 1996), LIGAND (Goto et al., 1998), SciFinder, ChemSpider (Ayers, 2012), and SPRESI which can assist in retrieving various chemical reaction equations, but also require a rather cumbersome usage process and a high learning cost.

Neural network and small language model-based chemistry AI systems. In recent times, there has been a surge in the development of neural network-driven algorithms for chemical prediction (Rappoport et al., 2014; Wei et al., 2016; Simm & Reiher, 2017; Fooshee et al., 2018; Xie et al., 2021; Schwaller et al., 2021; Meuwly, 2021), marking significant progress in the integration of AI technologies into chemistry. However, these methods are limited in their ability to address a broad spectrum of chemical tasks, focusing instead on specific categories or a narrow range of challenges within the field. Subsequently, there have been some efforts to apply small language models in the field of chemistry (Kuenneth & Ramprasad, 2023; Flam-Shepherd et al., 2022; Fabian et al., 2020; Edwards et al., 2021; Liu et al., 2021). Similarly, however, these methods are also only capable of addressing a subset of chemical tasks.

LLM-based chemistry AI systems. With LLMs showing immense potential in Ai4Sci, many studies have also begun to apply LLMs to the field of chemistry. Jablonka et al. (2024) design a predictive chemistry method based on GPT-3. Chemcrow (Bran et al., 2023) utilizes an LLM-based agent to autonomously plan and execute the syntheses of an insect repellent, three organocatalysts, and guides the discovery of a novel chromophore. Jablonka et al. (2023) explore the potential applications of LLMs for chemistry, including predicting properties of molecules and materials, as well as designing novel interfaces for tools. However, these efforts merely involve using prompt engineering to leverage existing LLMs, rather than training LLMs capable of addressing various chemical tasks. Although there are a few works that have trained LLMs based on chemical data (Zhang et al., 2024; Zhao et al., 2024), existing efforts still suffer from issues such as insufficient model size and limited multilingual capabilities.

3 METHODOLOGY

In this section, we first introduce the SMILES notion in chemistry. Following that, we describe our unified modeling approach for SMILES and natural language, along with corresponding instruction fine-tuning templates. Then, we introduce the sources of our training datasets.

3.1 SMILES NOTION

SMILES (Simplified Molecular Input Line Entry System) is a notation system used to represent chemical structures in a concise and human-readable format. It consists of a string of characters that represent atoms, bonds, and sometimes other molecular features. In SMILES notation:

- Atoms are represented by their elemental symbols (e.g., "C" for carbon, "H" for hydrogen, "O" for oxygen).
- Bonds between atoms are indicated by hyphens ("-") for single bonds, "=" for double bonds, "#" for triple bonds, and ":" for aromatic bonds.
- Parentheses "()" are used to group atoms and bonds to indicate branching or cyclic structures.
- Other features such as charges, isotopes, and stereochemistry can also be specified using additional symbols and conventions. [5]

For example, we present Thiamine (vitamin B1), showcasing its molecular formula, structure, and SMILES formula in Figure 1.

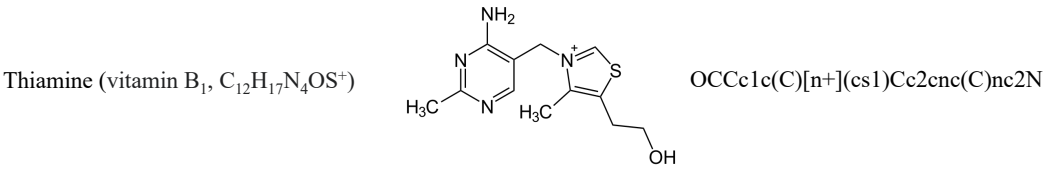

Figure 1: Example of molecule name, molecular formula, molecular structure, and its corresponding SMILES.

The SMILES formula captures all the atoms within the molecule and corresponds directly to its structural representation. The the direct conversion of SMILES formulas into corresponding molecular structures can be easily done by the tools like RDKit [6].

In other words, SMILES serves as a flexible sequence representation capable of capturing various attributes and structures, making it inherently suitable for sequence modeling with LLMs. Therefore, in this work, SMILES codes can be regarded as a distinct language, enabling the straightforward application of existing LLMs training methods.

3.2 UNIFIED MODELING

Viewing natural language as a specialized language, we can employ LLMs for unified modeling of natural language to SMILES, SMILES to natural language, SMILES to SMILES, and natural language to natural language. This naturally facilitates the completion of various chemistry tasks: **Molecule Description**, **Molecule design**, **Product Inference**, and **Retro-synthesis Prediction**. Additionally, we have also modeled the **Yield Prediction** task. We showcase our modeling approach in Figure 2. We model molecule description as bidirectional conversions between natural language and SMILES, as well as conversions between natural language. We model molecule design as a conversion from natural language to SMILES. We also model product inference and retro-synthesis prediction as conversions from SMILES to SMILES. Additionally, we have also included a task for yield prediction.

[5] More details can be found in `https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system`

[6] `https://rdkit.org/`

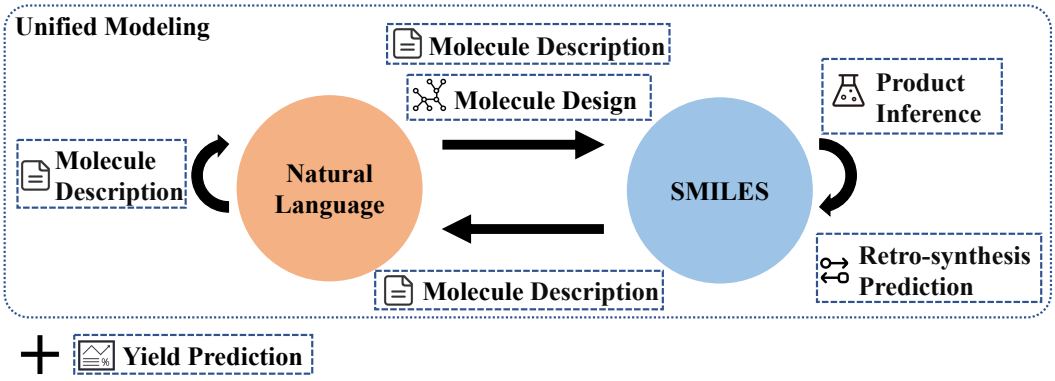

Figure 2: The illustration of our unified modeling between natural language and SMILES.

3.3 Chemistry Tasks and Prompt Templates

Following the modeling approach outlined above, we focus on the key tasks in the chemistry domain: **Molecule Description**, **Molecule design**, **Product Inference**, **Retro-synthesis Prediction**, and **Yield Prediction**. We construct instruction tuning datasets based on existing chemical, drug, and medicine datasets using prompt templates, in order to train models capable of addressing these tasks.

Retro-synthesis Prediction. Retro-synthesis Prediction is a crucial task for chemistry. It involves inferring possible reaction pathways and conditions by given product molecules, thereby reverse-predicting the synthetic route to generate the product. Retro-synthesis prediction enables researchers to explore and discover new organic molecular structures more rapidly, which is essential for fields such as organic synthesis chemistry and drug discovery. We train the model's retro-synthesis prediction capability using two subtasks: 1) Reactant and catalyst prediction: Given a product, predict the potential catalysts and reactants that may be required. 2) Reactant prediction: Given a product and catalyst, predict the reactants.

Product Inference. Product inference aims at predicting the products based on given starting materials and specific reaction conditions, which holds significant importance in fields such as organic synthesis and drug design. We train the model's product inference capability using two subtasks: 1) Product and catalyst prediction: Given reactants, predict the potential catalysts and products that may be involved. 2) Product prediction: Given products and catalysts, predict the reactants involved.

Molecule Design. Molecule Design is a field involving the creation of new molecules using theoretical and computational methods to produce molecular structures with specific properties or functionalities. This field plays a crucial role in various domains including chemistry, drug design, and materials science. The aim of molecule design is to systematically generate molecules with desired properties and activities to meet specific application needs. This work fully considers over a hundred molecular properties, such as molecule weight, valence electron count, Balaban J value, BertzCT value, number of heavy atoms, number of NHs or OHs, and number of nitrogen and oxygen atoms. It is hoped that the LLM can take into account researchers' specific requirements for molecule properties of catalysis, products, and reactants of chemical reactions. To train the model's molecular design capability, the following three tasks are adopted: 1) Specifying catalyst molecular properties: Given reactants to produce a specific product, the catalyst is required to meet certain properties. 2) Specifying reactant and catalyst molecular properties: Given the desired product to be synthesized, both reactants and catalysts are required to meet certain properties. 3) Specifying reactant, catalyst, and product properties: The model is required to provide a chemical reaction with specified molecular properties for reactants, catalysts, and products.

Molecule Description. Molecule description refers to using computational models to predict and describe the function, effects, and related properties of a molecule given its name, SMILES, or other representations. We adopt the following eight subtasks to train the model's molecule description capability. We not only utilize chemical data for training to enable the model to fully understand and perceive the correspondence between molecular names in both English and Chinese, molecule

descriptions, molecule SMILES, and molecule IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) names, thus obtaining strong molecular description capabilities, but also incorporate some pharmaceutical data to enhance the model's ability in the pharmaceutical field: 1) Given the molecular Chinese name, generate the English name and description. 2) Given the molecular English name, generate the Chinese name and description. 3) Given the molecular description, generate the Chinese name and English name of the molecule. 4) Given the molecular description, generate the IUPAC name and SMILES code of the molecule. 5) Given the molecular SMILES code, generate the IUPAC name and SMILES code of the molecule. 6) Given the molecular IUPAC name, generate the SMILES code and description of the molecule. 7) Given the Chinese name of a drug, generate the English name and description. 8) Given the English name of a drug, generate the Chinese name and description. 9) Given the description of a drug, generate the Chinese name and English name.

Yield Prediction. Yield prediction in chemical reactions refers to the estimation, through experimental or computational methods, of the ratio between the actual quantity of products generated in a chemical reaction and the theoretically maximum possible yield. We also train the model by predicting corresponding yields for given chemical reactions.

3.4 Data Source

We utilized publicly available high-quality datasets in the field of chemistry, as well as close-source datasets within our own team, as the raw datasets. Then, we transform them into instruction tuning datasets using the aforementioned prompt templates.

3.4.1 Publicly Available Datasets

- **USPTO** (Lowe, 2012) USPTO collects reaction data extracted through text mining from United States patents published between 1976 and September 2016.
- **CHEBI** (Degtyarenko et al., 2007) Chemical Entities of Biological Interest (CHEBI) is a freely available dictionary of molecular entities focused on "small" chemical compounds. The term "molecular entity" refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms.
- **CJHIF** (Jiang et al., 2021) Chemical Journals with High Impact factors (CJHIF) is a high-quality dataset containing a large number of chemical reaction equations extracted from various chemical journals.
- **PubChem** (Kim et al., 2016) PubChem is an open chemistry database at the National Institutes of Health (NIH), which mostly contains small molecules, but also larger molecules such as nucleotides, carbohydrates, lipids, peptides, and chemically-modified macromolecules.
- **Text2Mol** (Edwards et al., 2021) Text2Mol provides a large amount of data containing natural language descriptions of molecules.

3.4.2 Close Source Datasets

- **Drug Instruction** We collect a large number of drug names, drug descriptions, and corresponding molecular formulas from drug instruction to enhance the model's capabilities in the pharmaceutical domain.
- **Organic Compound Manual** We have a large collection of private organic compound manuals, containing information such as organic compound names, compound descriptions, compound SMILES, etc.
- **Molecular Formula and Name Reference Table** We have collected a large amount of publicly available data on compound names and their corresponding molecular formulas.
- **SMILES, IUPAC Names, and Molecular Descriptions Reference Table** We have collected data on SMILES, IUPAC names, and their corresponding molecular descriptions.

3.5 Data Transformation For Instruction Tuning

We extract reaction data into reactant SMILES, catalyst SMILES, product SMILES, and yield data. Then we conduct data augmentation, that is if there are multiple reactants, catalysts, or products, we shuffle the SMILES of these compounds.

For retro-synthetic prediction, product inference, and yield inference, we organize reactant SMILES, catalyst SMILES, product SMILES, and yield data according to the prompt templates. For molecule design, we use the RDKit tool to randomly select 1-20 properties from a candidate pool of 172 properties to fill in the prompt templates. We present partial prompt templates and examples corresponding to different chemical tasks in Appendix A and Appendix B.

3.6 Data Details

Task	Amount
Retro-synthesis Prediction	30114006
Product Inference	30114006
Molecule Design	40695857
Molecule Description	210469
Yield Prediction	10775991
Total	111910329

Table 1: Data details.

Table 1 lists the data scale used for each task. We have over a hundred million data entries in total, with an average length exceeding 150 tokens. The total number of tokens trained exceeds 15 billion.

4 Training Details

4.1 Base Model

We select our team's self-developed BatGPT-15B model (Li et al., 2023c) as the base model for instruction tuning. BatGPT-15B is a large bilingual model for both Chinese and English, pre-trained using bidirectional autoregressive methods, and has demonstrated excellent performance on public benchmarks such as CMMLU (Li et al., 2023a).

4.2 Vocabulary Expansion

Since the BatGPT-15B model is originally designed for natural language, particularly Chinese and English, it lacks comprehensive coverage of specialized terms in chemistry or SMILES. Consequently, expanding its vocabulary becomes necessary. We employ the Byte Pair Encoding (BPE) algorithm to train a vocabulary using diverse training data, encompassing various forms of molecular SMILES, chemical equation SMILES expressions, molecular names, and more. We also includ all chemical element symbols in the augmented vocabulary to empower the model with the potential to handle all chemical elements. Subsequently, we merge this augmented vocabulary with that of BatGPT-15B, ultimately yielding a final vocabulary size of 151851.

4.3 Training Settings

We train our model using the deepspeed zero2 strategy on an Nvidia A800 GPU cluster. We set the maximum length to 2048, the batch size per GPU to 8, utilize the AdamW optimizer with a learning rate of 2e-4, and employ the cosine learning rate schedule strategy. We enable gradient checkpointing and set max gradient normalization to 1.0 and weight decay to 0.1.

5 EXPERIMENTS

5.1 SETTINGS

We primarily evaluate BatGPT-15B in the field of predictive retrosynthesis, which is of utmost interest in current chemical engineering. We evaluate the model using the retro-synthesis prediction task from USPTO-50k and BioChem (Zheng et al., 2022), wherein the model is tasked with predicting the corresponding reactants and catalysts given the products.

We evaluate the model performance from two aspects: 1. Coverage, which indicates how much of the true products are covered by the model outputs, reflecting the accuracy of the reaction predictions. Due to the inability of LLMs to guarantee that the predicted SMILES codes are entirely consistent with the ground truth, there might be cases where different SMILES codes are predicted for the same molecule. Therefore, we do not opt for the traditional Exact Match metric. 2. Validity, which measures how many of the SMILES codes predicted by the model are legal and do not violate chemical principles. We also conduct multiple experiments by adjusting the top-k hyperparameter of the model inference.

5.2 MAIN RESULTS

Dataset	Top-k	Coverage	Validity
USPTO-50k	1	49.17%	99.98%
	3	56.11%	99.97%
	5	63.71%	99.98%
	10	69.15%	99.97%
BioChem	1	25.79%	99.82%
	3	31.81%	99.73%
	5	34.62%	99.75%
	10	38.08%	99.77%

Table 2: The main result of retro-synthesis prediction.

We present the main results of inverse synthesis prediction in Table 2. For coverage, setting a larger top-k will result in greater coverage. The maximum coverage achieved on USPTO-50k and BioChem reach 69.15% and 38.08%, which are comparable to the previous state-of-the-art results. Regarding validity, BatGPT-Chem surpasses 99.77% in all predictions, indicating that the model, after extensive training, almost never outputs SMILES codes that violate chemical rules. In addition to the aforementioned benchmarks, we also conduct experiments on a private benchmark dataset. We showcase some cases in the Appendix C.

6 CONCLUSION

LLMs have made remarkable strides across diverse domains and possess the potential to drive advancements in Ai4Sci. With their adeptness in learning from sequential data, LLMs are ideally suited for the chemical domain, where common representations like SMILES are also sequential data, thus naturally aligning with the learning capabilities of LLMs. Thus in this work, we introduce BatGPT-Chem, a foundational large model for chemical engineering. We unify chemistry tasks with natural language and SMILES, design prompt templates, and generate instruction tuning data. We then train BatGPT-15B with over 100M instruction tuning data, empowering it to handle tasks like Molecule Description, Molecule Design, Retro-synthesis Prediction, Product Inference, and Yield Prediction.

REFERENCES

Meredith Ayers. Chemspider: the free chemical database. *Reference reviews*, 26(7):45–46, 2012.

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

Andres M Bran, Sam Cox, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

Elias James Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969.

Randall Davis and Jonathan J King. The origin of rule-based systems in ai. *Rule-based expert systems: The MYCIN experiments of the Stanford Heuristic Programming Project*, 1984.

Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. *Nucleic acids research*, 36 (suppl_1):D344–D350, 2007.

James Dugundji and Ivar Ugi. An algebraic model of constitutional chemistry as a basis for chemical computer programs. In *Computers in chemistry*, pp. 19–64. Springer, 2006.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. Molecular representation learning with language models and domain-relevant auxiliary tasks, 2020.

Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293, 2022.

David A Fletcher, Robert F McMeeking, and Donald Parkin. The united kingdom chemical database service. *Journal of chemical information and computer sciences*, 36(4):746–749, 1996.

David Fooshee, Aaron Mood, Eugene Gutman, Mohammadamin Tavakoli, Gregor Urban, Frances Liu, Nancy Huynh, David Van Vranken, and Pierre Baldi. Deep learning for chemical reaction prediction. *Molecular Systems Design & Engineering*, 3(3):442–452, 2018.

Susumu Goto, Takaaki Nishioka, and Minoru Kanehisa. Ligand: chemical database for enzyme reactions. *Bioinformatics (Oxford, England)*, 14(7):591–599, 1998.

Kevin Maik Jablonka, Qianxiang Ai, Alexander Al-Feghali, Shruti Badhwar, Joshua D Bocarsly, Andres M Bran, Stefan Bringuier, L Catherine Brinson, Kamal Choudhary, Defne Circi, et al. 14 examples of how llms can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2(5):1233–1250, 2023.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, pp. 1–9, 2024.

Shu Jiang, Zhuosheng Zhang, Hai Zhao, Jiangtong Li, Yang Yang, Bao-Liang Lu, and Ning Xia. When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. *IEEE Access*, 9:85071–85083, 2021.

Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

Christopher Kuenneth and Rampi Ramprasad. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature Communications*, 14(1):4099, 2023.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023a.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *arXiv preprint arXiv:2306.06615*, 2023b.

Zuchao Li, Shitou Zhang, Hai Zhao, Yifei Yang, and Dongjie Yang. Batgpt: A bidirectional autoregessive talker from generative pre-trained transformer. *arXiv preprint arXiv:2307.00360*, 2023c.

Zhichao Liu, Ruth A Roberts, Madhu Lal-Nag, Xi Chen, Ruili Huang, and Weida Tong. Ai-based language models powering drug discovery and development. *Drug Discovery Today*, 26(11): 2593–2607, 2021.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, 2012.

Gang Lu, Sufen Ai, and Junbai Li. Layer-by-layer assembly of human serum albumin and phospholipid nanotubes based on a template. *Langmuir*, 21(5):1679–1682, 2005.

Markus Meuwly. Machine learning for chemical reactions. *Chemical Reviews*, 121(16):10218–10239, 2021.

Chen Qian, Huayi Tang, Zhirui Yang, Hong Liang, and Yong Liu. Can large language models empower molecular property prediction? *arXiv preprint arXiv:2307.07443*, 2023.

Dmitrij Rappoport, Cooper J Galvin, Dmitry Yu Zubarev, and Alán Aspuru-Guzik. Complex chemical reaction networks from heuristics-aided quantum chemistry. *Journal of chemical theory and computation*, 10(3):897–907, 2014.

Timothy D Salatin and William L Jorgensen. Computer-assisted mechanistic evaluation of organic reactions. 1. overview. *The Journal of Organic Chemistry*, 45(11):2043–2051, 1980.

Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology*, 2(1):015016, 2021.

Gregor N Simm and Markus Reiher. Context-driven exploration of complex chemical reaction networks. *Journal of chemical theory and computation*, 13(12):6108–6119, 2017.

Stanley G Smith and Bruce Arne Sherwood. Educational uses of the plato computer system: The plato system is used for instruction, scientific research, and communications. *Science*, 192(4237): 344–352, 1976.

Jennifer N Wei, David Duvenaud, and Alán Aspuru-Guzik. Neural networks for the prediction of organic chemistry reactions. *ACS central science*, 2(10):725–732, 2016.

Xiaowei Xie, Evan Walter Clark Spotte-Smith, Mingjian Wen, Hetal D Patel, Samuel M Blau, and Kristin A Persson. Data-driven prediction of formation mechanisms of lithium ethylene monocarbonate with an automated reaction network. *Journal of the American Chemical Society*, 143 (33):13245–13258, 2021.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*, 2024.

Zihan Zhao, Da Ma, Lu Chen, Liangtai Sun, Zihao Li, Hongshen Xu, Zichen Zhu, Su Zhu, Shuai Fan, Guodong Shen, et al. Chemdfm: Dialogue foundation model for chemistry. *arXiv preprint arXiv:2401.14818*, 2024.

Shuangjia Zheng, Tao Zeng, Chengtao Li, Binghong Chen, Connor W Coley, Yuedong Yang, and Ruibo Wu. Deep learning driven biosynthetic pathways navigation for natural products with bionavi-np. *Nature Communications*, 13(1):3342, 2022.

A PROMPT TEMPLATES

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Retrosynthesis Prediction	Reactant and Catalyst Prediction	1. 给定产物的SMILES码为{}，对应的反应物和催化剂可能为： 2. 已知产物的SMILES码为{}，相应的反应物和催化剂可能为： 3. 给定产物的SMILES码为{}，可以推断出潜在的反应物和催化剂为： 4. 对于SMILES码为{}的产物，相应的反应物和催化剂可能为： 5. 现有产物，其SMILES码为{}，则用于合成它们的反应物和催化剂可能为：	1. Given the SMILES codes of the products , the corresponding reactants and the catalysts can be: 2. Knowing the SMILES codes of the products, the corresponding reactants and catalysts can be: 3. Given the SMILES codes {} of the products, potential reactants and catalysts can be deduced as: 4. With the SMILES codes of the products provided, potential reactants and catalysts can be deduced as follows: 5. Given the SMILES codes {} of the products, potential reactants and catalysts can be inferred as:
	Reactant Prediction	1. 给定产物的SMILES码{}和催化剂的SMILES码{}，可能的反应物包含： 2.给定产物的SMILES码{}和催化剂的SMILES码{}，潜在的反应物可能包括： 3.当一个化学反应的产物的SMILES码为{}和催化剂的SMILES码为{}时，可能的反应物有： 4.当反应得到的产物的SMILES码为{}，催化剂为{}，可能的反应物包含： 5.对于SMILES码为{}的产物和SMILES码为{}的催化剂，这个反应可能的反应物是：	1. Given the SMILES codes of the products {} and the catalysts {}, the possible reactants can be: 2. Provided with the SMILES codes of the products {} and the catalysts {}, potential reactants may include: 3. When given the SMILES codes of the products {} and the catalysts {}, the potential reactants could be: 4. When provided with the SMILES codes of the products {} and the catalysts {}, potential reactants to consider are: 5. With the SMILES codes of the products {} and the catalysts {} provided, potential reactants may encompass:

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Product Inference	Product and Catalyst Prediction	1. 给定反应物的SMILES码为{}，可以与之搭配的催化剂以及反应得到的产物为： 2. 现有反应物的SMILES码包括{}。可以与它们搭配的催化剂以及得到的产物为： 3. 对于SMILES码为{}的反应物，可以在反应时加入的催化剂以及得到的产物为： 4. 使用的反应物的SMILES码为{}，可以加入的催化剂和得到的产物可能为： 5. 对于反应物{}，潜在的可以加入的催化剂以及对应得到的产物可能为：	1. The SMILES codes of the reactants are {}. The corresponding catalysts it can pair with and the resulting products are: 2. The existing reactants include {}. The catalysts that can be matched with them and the resulting products are as follows: 3. The given reactants are {}. The catalysts that can be combined with them, along with the resulting products, are as follows: 4. With the SMILES codes of the products {} provided, potential reactants and catalysts can be deduced as follows: 5. With the specified reactants as {}, the associated catalysts and the resulting products can be:
	Product Prediction	1. 给定反应物的SMILES码为{}，催化剂的SMILES码为{}，产物为： 2. 反应物的SMILES码为{}，使用的催化剂的SMILES码为{}，则反应产生的产物为： 3. 当给定的反应物的SMILES码为{}，催化剂的SMILES编码为{}时，产物为： 4. 当反应物{}加入催化剂{}进行反应的时候，得到的产物为： 5.{}表示反应物的SMILES编码，{}表示催化剂，反应得到的产物为：	1. Given the SMILES codes of the reactants {}, the catalysts {}, the products are: 2. With the SMILES codes of the reactants {} and the catalysts {}, the resulting products are: 3. When the SMILES codes of the reactants {} and the catalysts {} are given, the products are: 4. The reactants {} and the catalysts {} will determine the resulting products: 5. The SMILES codes of the reactants are {}, the catalysts are {}, and the products are:

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Molecular Design	Specify Catalyst Molecular Properties	1. 要求用反应物{}制备产物{}，要求催化剂满足条件：{}，催化剂可以是：	1. Given reactants {} and products {}, the catalysts are required that meets the conditions: {}, so the catalysts can be:
		2. 实现由反应物{}合成产物{}的过程需要考虑满足条件：{}的催化剂，催化剂可以是：	2. Provided with reactants {} and products {}, the catalysts needed must satisfy the conditions: {}. Possible catalysts include:
		3. 由反应物{}合成产物{}的过程，选用的催化剂需要满足条件：{}，催化剂可以被选择为：	3. Given reactants {} and products {}, the catalysts required should meet the specified conditions: {}. Potential catalysts may be:
		4. 为了使反应物{}合成产物{}，可选择一种满足条件：{}的催化剂，催化剂可以是：	4. With specified reactants {} and products {}, the catalysts needed can meet the conditions: {}. Potential catalysts can be:
		5. 用反应物{}制备产物{}，可以添加催化剂，要求其符合条件：{}，催化剂可为：	5. Given the specified reactants {} and products {}, the catalysts required have the capability to satisfy the conditions: {}. Possible catalysts include:
	Specify Reactant and Catalyst Molecular Properties	1. 为了制备产物{}，要求反应物满足条件：{}，催化剂满足条件：{}，反应物和催化剂分别可以是：	1. To synthesize the products {}, it is required that the reactants meet the conditions: {}, and the catalysts satisfy the conditions: {}. The reactants and the catalyst can be:
		2. 要合成产物{}，反应物要满足：{}的条件，而催化剂也需要满足：{}的要求，反应物和催化剂分别可以是：	2. To synthesize the products {}, the reactants need to meet the conditions: {}, and the catalysts also need to satisfy the requirements: {}. The reactants and the catalysts can be:
		3. 实现产物{}的合成过程需要考虑满足条件：{}的反应物，以及满足条件：{}的催化剂，则反应物和催化剂可以是：	3. The synthesis process for the products {} involves considering reactants that meet the conditions: {}, as well as the catalysts that satisfy the requirements: {}. The reactants and the catalysts can be::
		4. 为了合成产物{}，要求反应物满足:{}的条件，催化剂满足：{}的条件，则可以选择的反应物和催化剂分别是：	4. To synthesize the products {}, it is required that the reactants meet the conditions: {}, and the catalysts satisfy the conditions: {}. The possible choices for reactants and catalysts are:
		5.要合成产物{}，选用的反应物需要满足条件：{}，催化剂需要满足条件：{}，则可以选择：	5.To synthesize the products {}, the chosen reactants need to meet the conditions: {}, and the catalysts should satisfy the conditions: {}. The selection can include:

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Molecular Design	Specify Reactant, Catalyst and Product Molecular Properties	1. 对于一个可以进行的化学反应，要求反应物满足条件：{}，催化剂满足条件：{}，产物满足条件：{}，则这个反应的反应物、催化剂和产物分别可以是：	1.For a feasible chemical reaction, the reactants meet the conditions: {}, the catalysts satisfy the conditions: {}, and the products fulfill the conditions: {}. The specific reactants, catalysts, and products for this reaction can be:
		2. 对于一个可发生的化学反应，要求反应物符合条件：{}，同时催化剂要满足条件：{}，产物也需要符合条件：{}。这个反应的具体反应物、催化剂和产物分别可以是：	2. For a possible chemical reaction, it is required that the reactants meet the conditions: {}, and the catalysts satisfy the conditions: {}, while the products also fulfill the conditions: {}. The specific reactants, catalysts, and products for this reaction can be:
		3. 进行一种可行的化学反应时，要求反应物满足条件：{}，催化剂符合条件：{}，产物符合条件：{}。这个反应所涉及的反应物、催化剂和产物分别可以是：	3. When conducting a feasible chemical reaction, it is required that the reactants meet the conditions: {}, the catalysts should satisfy the conditions: {}, and the products comply with the conditions: {}. The specific reactants, catalysts, and products involved in this reaction can be:
		4. 在进行某一可实施的化学反应时，反应物的选择需要满足条件：{}，催化剂也需要符合条件：{}，得到的产物满足条件：{}。这个反应中的具体反应物、催化剂和产物分别可以是：	4. When conducting a feasible chemical reaction, the choice of reactants needs to meet the conditions: {}, the catalysts also satisfy the conditions: {}, and the resulting products fulfill the conditions: {}. The specific reactants, catalyst, and product involved in this reaction can be:
		5. 对于一种可进行的化学反应，反应物满足条件：{}，催化剂满足条件：{}，产物满足条件：{}。则所选的反应物、催化剂和对应产物可以是：	5. For a possible chemical reaction, the reactants meet the conditions: {}, the catalysts meet the conditions: {}, and the products comply with the conditions: {}. The chosen reactants, catalysts, and the corresponding products can be:

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Molecular Description	Given the Chinese name of the drug, generate an English name and description	1. 给定一个药物的中文名称{}，它的对应英文名称和一些相关描述是： 2. {}这种药物对应的英文名和一些相关描述如下所示： 3. 对于{}这一药物，其英文名和一些药物描述如下： 4. 以下是关于{}药物对应的英文名称和一些药物描述： 5.一个药物的中文名称是{}，它的英文名和药物描述是：	
	Given the English name of the drug, generate an Chinese name and description	1. 英文名为{}的药物的中文名和药物描述如下： 2. 英文名为{}的药物对应的中文名和一些相关描述为： 3. 给出{}这种药物的中文名和一系列相关的药物描述： 4. 以下是关于{}药物对应的中文名称和一些药物描述： 5.一个药物的英文名称是{}，它的中文名和药物描述是：	
	Given the drug description, generate Chinese and English names	1. 药物描述为{}对应的药物中文名和英文名分别是： 2. 一个药物的描述为{}，那么它可能的中文名称和英文名称分别是： 3. 如果药物的描述为{}，那么可能的中文名称和英文名称是： 4.{}这个描述对应的药物的中文名和英文名称可能是： 5. 如果有一个药物被描述为{}，那么它的可能中文名称和英文分别是：	
	Given the molecular name, generate the molecular formula	1. 给定一个分子名{}，它对应的分子式为： 2. 一个分子的名字是{}，其相应的分子式是： 3. 以{}为名字的分子，其分子式是： 4. 对于一个以{}为名字的分子，其分子式是： 5. {}所代表的分子的分子式是：	
	Given the molecular formula, generate the molecular name	1. 一个分子的分子式是{}，它的名字是： 2. 分子式为{}的分子对应的名字是： 3. {}分子式对应的分子名是： 4. 对于分子式{}，它的名字为： 5. {}分子式的名称是：	

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Molecular Description	Given the molecular description, generate molecular IUPAC name and SMILES code		1. Given a description of a molecule: {}, the possible IUPAC name and corresponding SMILES code for this molecule are: 2. When provided with a molecule description: {}, the potential IUPAC name and corresponding SMILES code for the molecule can be determined as: 3. In the context of a molecule description: {}, the molecule's potential IUPAC name and its corresponding SMILES code are: 4. In the case of a molecule described as: {}, the molecule's possible IUPAC name and the corresponding SMILES code can be: 5. A molecule described as: {} may have a potential IUPAC name and corresponding SMILES code:
	Given the molecular SMILES code, generate molecular IUPAC name and description		1. Given a SMILES code of a molecule: {}, the possible IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name and corresponding description for this molecule are: 2. If given a SMILES code representing a molecule as {}, the potential IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name and related description can be: 3. In the case of a molecule with the SMILES code {}, the potential IUPAC name and description for the molecular are: 4. When provided with a SMILES code {} for a molecule, the IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name and associated description for the compound can be: 5. Given a SMILES code {} for a molecule, the potential IUPAC name and its related description can be identified as:

Chemical Task	Subtask	Chinese Prompt Template	English Prompt Template
Molecular Description	Given the molecular IUPAC name, generate molecular description and SMILES code		1. Given a IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name of a molecule: {}, the possible SMILES code and corresponding description for this molecule are: 2. The SMILES code and description for the molecule with the IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name {} are: 3. Providing the IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name {} for a molecule, the possible SMILES code and accompanying description are: . Given the IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name {} for a molecule, the potential SMILES code and its corresponding description are: 5. Providing the IUPAC name {} for a molecule, the potential SMILES code and a corresponding description can be:
Yield Prediction		1. 对于一个化学反应，其反应物的SMILES码 {}，催化剂 {}，产物 {}，期望的产率值是： 2. 对于一个反应物的SMILES码{}，催化剂{}，产物{}的化学反应，期望的产率值为： 3. 一个化学反应的反应物、催化剂和产物的SMILES码分别为{}、{}、{}，其期望产率是 4. 为了使反应物{}合成产物{}，可选择一种满足条件：{}的催化剂，催化剂可以是： 5. 对于一个反应物的SMILES码为 {}、催化剂的SMILES码为 {}、产物的SMILES码为 {} 的化学反应，其期望产率是：	1. Given the SMILES codes of the reactants {}, the catalysts {}, the products {}, the expected yield value of this chemical reaction is: 2. By examining the SMILES codes of the reactants {}, the catalysts {}, and the products {}, the expected yield value of this chemical reaction can be estimated as: 3. Through an examination of the SMILES codes of the reactants {}, the catalysts {}, and the products {}, the expected yield value of this chemical reaction is: 4. The reactants, catalysts, and products of a chemical reaction are {}, {} and {}, and its expected yield is: 5. The reactants {}, catalysts {}, and products {} define a chemical reaction, and its expected yield is:

B PROMPT EXAMPLE

Chemical Task	Subtask	Chinese Prompt	English Prompt
Retrosynthesis Prediction	Reactant and Catalyst Prediction	给定产物的SMILES码为COc1ccccc1OCCN1CC(COc2ccc3c(c2[nH]c2ccccc32)OCC1=O，可以推断出潜在的反应物和催化剂为：	Given the product SMILES codes COc1ccccc1OCCN1CC(COc2ccc3c(c2)[nH]c2ccccc32)OCC1=O, the corresponding reactants and catalysts can be:
	Reactant Prediction	1. 给定产物的SMILES码CC(=C)c1cccc(C)c1Br和催化剂的SMILES码[K+], CC(C)(C)[O-]和C1CCOC1，可能的反应物包含： 2.（不指定催化剂）给定产物的SMILES码CC(=C)c1cccc(C)c1Br，潜在的反应物可能包括：	1. Given the SMILES codes of the products CC(=C)c1cccc(C) c1Br and the catalysts [K+], CC(C) (C)[O-] and C1CCOC1, the possible reactants can be: 2. (No catalyst specified) Provided with the SMILES codes of the products CC(=C)c1cccc(C)c1Br, potential reactants may include:
Product Inference	Product and Catalyst Prediction	给定反应物的SMILES码为C(c1ccccc1)Br, c1(C)ccccc1, C(C)(=O)C 和 C=C，可以与之搭配的催化剂以及反应得到的产物为：	The SMILES codes of the reactants are C(c1ccccc1)Br, c1(C)ccccc1, C(C)(=O)C and C=C. The corresponding catalysts it can pair with and the resulting products are:
	Product Prediction	1. 给定反应物的SMILES码为O=CN,O(C(=O)C(OC(C)=O)N1C(CC#C[Si](C)(C)C)CC1=O)C(C)C, C=C和COC(C)=O，催化剂的SMILES码为C(Cl)Cl和Cl[Sn](Cl)(Cl)Cl，产物为: 2.（不指定催化剂）反应物的SMILES码为c1c(F)cc(F)c(N)c1和c1(CCCCn2nncc2)ccc(cc1)OCc1occ(C(O)=O)n1，则反应产生的产物为：	1. Given the reactant SMILES codes O=CN, O(C(=O)C(OC(C)=O)N1C(CCC#C[Si](C)(C)C)CC1=O)C(C)C, C=C and COC(C)=O, the catalysts C(Cl)Cl and Cl[Sn](Cl)(Cl)Cl, the products are: 2. (No catalyst specified) With the reactant SMILES codes Cl, C1OCCOC1 and n1c2c([nH]c1-c1c(I)ccnc1OC)cc(C#N)cc2C, the resulting products are:
Molecular Design	Specify Catalyst Molecular Properties	用反应物 [Na+], c1(C(c2ccc(O)cc2)=O)ccccc1, [H-], C(Br)C(CO)(CO)CBr 和 C(Br)C1(CO)COC1 制备产物 O=C(C1=[CH][CH]=[CH][CH]=[CH]1)C1=[CH][CH]=C(O[CH2]C2([CH2][OH])[CH2]O[CH2]2)[CH]=[CH]1，可以添加催化剂，要求其符合条件：杂原子的数量≥ 0.0 并且< 2.4, 酰胺的数量≥ 0.0 并且<0.7, 氢键受体的数量≥ 0.0 并且 < 1.4，催化剂可为：	Given reactants CO, COC, NC=O and COc1ccccc1OCCN(CC(O)COc1ccc2c(c1)[nH]c1ccccc21)C(=O)CCl and products COc1ccccc1OCCN1CC(COc2ccc3c(c2)[nH]c2ccccc32)OCC1=O, the catalysts required should meet the specified conditions: the number of Heteroatoms ≥ 0.0 and < 2.4, the number of Hydrogen Bond Donors ≥ 0.0 and < 1.9, the total number of NHs or OHs ≥ 0.0 and < 2.5, the number of Hydrogen Bond Acceptors ≥ 0.0 and < 1.4, Wildman-Crippen LogP value ≥ -4.1 and < -0.1, the exact molecular weight of the molecule ≥ 0.0 and < 204.6, Wildman-Crippen MR value ≥ 0.0 and < 52.1. Potential catalysts may be:

Chemical Task	Subtask	Chinese Prompt	English Prompt
Molecular Design	Specify Reactant and Catalyst Molecular Properties	要合成产物[CH3]C([CH3])([CH3])OC(=O)N1[CH2][CH]=C(N([CH2]C2=[CH][CH]=[CH][CH]=[CH]2)C(=O)C2=C(I)[CH]=[CH][CH]=[CH]2)[CH2][CH2]1，反应物要满足：分子的价电子数≥179.0 并且<286.8，分子的NH和OH总数量≥2.5 并且<5.5，卤素原子的数量≥1.0 并且<3.3，旋转键的数量≥0.0 并且 <6.0，羰基氧原子的数量≥ 1.6 并且 <3.6，Wildman-Crippen MR值≥129.0 并且 <205.9 的条件，而催化剂也需要满足：分子的准确分子量≥0.0 并且 <204.6，卤素原子的数量≥0.0 并且 <1.0，氢键供体的数量≥0.0 并且 <1.9，分子的氮氧原子总数量≥0.0 并且<1.3 的要求，反应物和催化剂分别可以是：	To synthesize the products [CH3]C([CH3])([CH3])OC(=O)N1[CH2][CH]=C(N([CH2]C2=[CH][CH]=[CH][CH]=[CH]2)C(=O)C2=C(I)[CH]=[CH][CH]=[CH]2)[CH2][CH2]1, the reactants need meet the conditions: the number of amides ≥0.7 and <2.1, the number of ether oxygens (including phenoxy) ≥0.0 and <1.5, the number of Hydrogen Bond Acceptors ≥ 1.4 and <6.1, the total number of NHs or OHs ≥2.5 and <5.5, the number of Hydrogen Bond Donors ≥0.0 and <1.9, Balaban's J value ≥-0.4 and <0.8, the number of heavy atoms ≥32.7 and <52.5, and the catalysts also need to satisfy the requirements: Balaban's J value ≥1.9 and <3.1, the number of Hydrogen Bond Acceptors ≥0.0 and <1.4, the number of benzene rings ≥0.0 and <1.6. The reactants and the catalysts can be:
	Specify Reactant, Catalyst and Product Molecular Properties	进行一种可行的化学反应时，要求反应物满足条件：Wildman-Crippen MR值≥ 129.0 并且 < 205.9，分子的价电子数≥ 179.0 并且 < 286.8，醚氧原子的数量（包括苯氧基）≥ 0.0 并且 < 1.5，BertzCT值≥ 146.5 并且 < 925.1，氢键供体的数量≥ 0.0 并且 < 1.9，Balaban's J值≥ -0.4 并且 < 0.8，分子的准确分子量≥ 493.6 并且 < 782.5，酰胺的数量≥ 0.7 并且 < 2.1，分子的NH和OH总数量≥ 2.5 并且 < 5.5，氢键受体的数量≥ 1.4 并且 < 6.1，催化剂符合条件：Wildman-Crippen LogP值≥ -0.1 并且 ¡< 3.8，Balaban's J值≥ 1.9 并且 < 3.1，Wildman-Crippen MR值≥ 0.0 并且 < 52.1，分子的重原子数≥ 0.0 并且 < 13.0，酰胺的数量≥ 0.0 并且 < 0.7，分子的准确分子量≥ 0.0 并且 < 204.6，氢键受体的数量≥ 0.0 并且 <1.4，氢键供体的数量≥ 0.0 并且 < 1.9，分子的氮氧原子总数量≥ 0.0 并且 < 1.3，醚氧原子的数量（包括苯氧基）≥ 0.0 并且 < 1.5，产物符合条件：氢键受体的数量≥ 1.4 并且 < 6.1，酰胺的数量≥ 0.7 并且 < 2.1，苯环的数量≥ 1.6 并且 < 3.3，分子的准确分子量≥ 493.6 并且 < 782.5，Wildman-Crippen MR值≥ 52.1 并且 < 129.0。这个反应所涉及的反应物、催化剂和产物分别可以是：	When conducting a feasible chemical reaction, it is required that the reactants meet the conditions: the number of ether oxygens (including phenoxy) ≥0.0 and <1.5, the number of heavy atoms ≥32.7 and <52.5, the total number of Nitrogens and Oxygens ≥1.3 and <7.1, the number of valence electrons the molecule ≥179.0 and <286.8, the total number of NHs or OHs ≥2.5 and <5.5, the number of Hydrogen Bond Acceptors ≥1.4 and <6.1, the catalysts should satisfy the conditions: the number of Hydrogen Bond Acceptors ≥0.0 and <1.4, Balaban's J value ≥1.9 and <3.1, Wildman-Crippen MR value ≥0.0 and <52.1, the number of carbonyl O ≥0.0 and <1.6, the number of valence electrons the molecule ≥ 0.0 and <71.3, the total number of Nitrogens and Oxygens ≥0.0 and <1.3, the number of amides ≥0.0 and <0.7, Wildman-Crippen LogP value ≥-0.1 and <3.8, the number of halogens atoms ≥0.0 and <1.0, the number of Heteroatoms ≥0.0 and <2.4, and the products comply with the conditions: the number of amides ≥0.7 and <2.1, the total number of NHs or OHs ≥0.0 and <2.5, the exact molecular weight of the molecule ≥493.6 and <782.5. The specific reactants, catalysts, and products involved in this reaction can be:

Chemical Task	Subtask	Chinese Prompt	English Prompt
	Given the drug Chinese name, generate English name and description	给定一个药物的中文名:先锋哌唑酮，它的对应英文名称和一些相关描述是	
	Given the drug English name, generate Chinese name and description	一个药物的英文名称是 Ped-el，它的中文名和药物描述是:	
	Given the drug description, generate Chinese and English names	一个药物的描述为 [类别]镇静催眠抗惊厥药,[适应症] 用于治疗神经衰弱、忆病、神经性失眠、精神兴奋状态。 [用量用法] 口服:每次10ml,1日3次。 [注意事项] 1.不宜用于浮肿和少尿及癫痫病人。 2.其他参见溴化钾及溴化铵。 [规格] 溶液:含溴化钾3%、溴化钠3%、溴化铵3%。，那么它可能的中文名称和英文名称分别是:	
Molecular Description	Given the molecular name, generate its molecular formula	给定一个分子名 trideuteriomethyl 2,2,2-tribromoacetate，它对应的分子式为:	
	Given the molecular formula, generate its molecular name	分子式为 C21H14K4N2O9的分子对应的名字是:	
	Given the molecular description, generate molecular IUPAC name and SMILES code		In the context of a molecule description: 'It has a role as an antimanic drug. It is an inorganic chloride and a lithium salt.', the molecule's potential IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name and its corresponding SMILES code are:
	Given the molecular SMILES code, generate molecular IUPAC name and description		Given a SMILES code '[Cl-].[K+]' for a molecule, the potential IUPAC name and its related description can be identified as:
	Given the molecular IUPAC name, generate molecular description and SMILES code		Given a IUPAC (International Union of Pure and Applied Chemistry chemical nomenclature) name of a molecule: 'magnesium;dichloride;hydrate', the possible SMILES code and corresponding description for this molecule are:
Yield Prediction		对于一个化学反应，其反应物的SMILES码 CCl.CBr.Brc1ccc(cc1)C(=O)CN(=O)=O.ON=C(Cl)c1cccc(Cl)c1，催化剂 CCN(CC)CC.CO，产物 Clc1cccc(c1)c1onc(c1N(= O)=O)-c1ccc(Br)cc1，期望的产率值是:	Given the molecular formulas of the reactants CCl.CBr.Brc1ccc(cc1)C(=O)CN(=O)=O.ON=C(Cl)c1cccc(Cl)c1, the catalysts CCN(CC)CC.CO, the products Clc1cccc(c1)-c1onc(c1N(=O)=O)-c1ccc(Br)cc1, the expected yield value is:

C Case Study

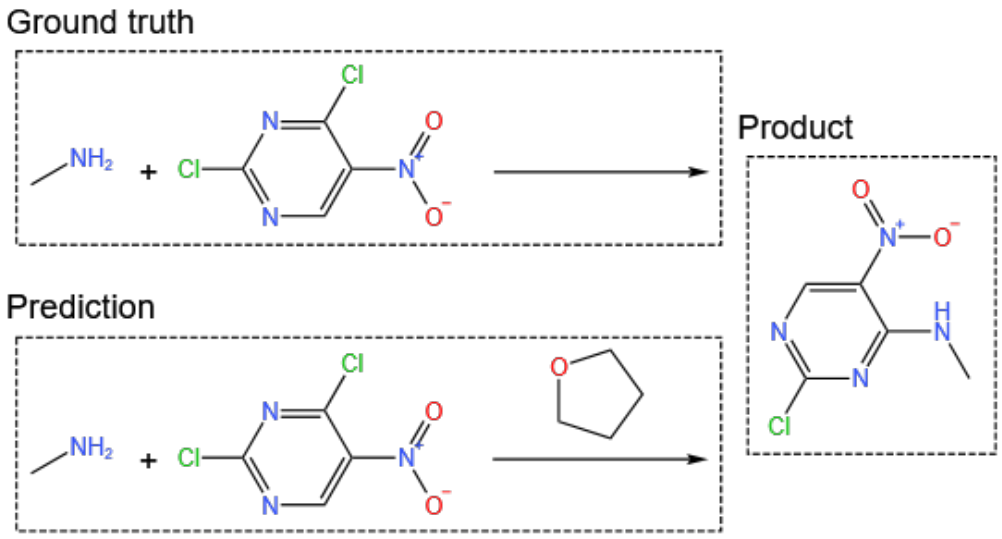

Figure 3: Case 1 from USPTO-50k. Give the product **CNc1nc(Cl)ncc1[N+](=O)[O-]**, the model successfully predicts the correct reactant **CN.O=[N+]([O-])c1cnc(Cl)nc1Cl**. It also simultaneously provides a potential catalyst **C1COCC1**, which is a commonly used catalyst.

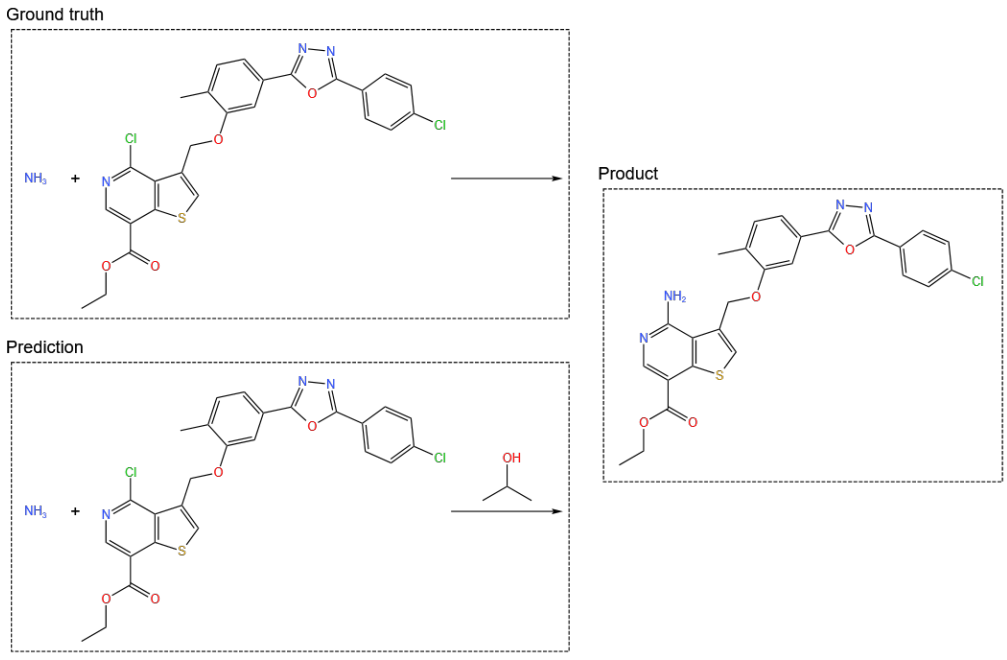

Figure 4: Case 2 from USPTO-50k. Give the product **CCOC(=O)c1cnc(N)c2c(COc3cc(-c4nnc(-c5ccc(Cl)cc5)o4)ccc3C)csc12**, the model successfully predicts the correct reactant **N.Clc1c2c(scc2COc2c(C)ccc(-c3nnc(-c4ccc(Cl)cc4)o3)c2)c(C(OCC)=O)cn1**. The model also predicts a catalyst **C(C)(O)C**, which could possibly act as a solvent.

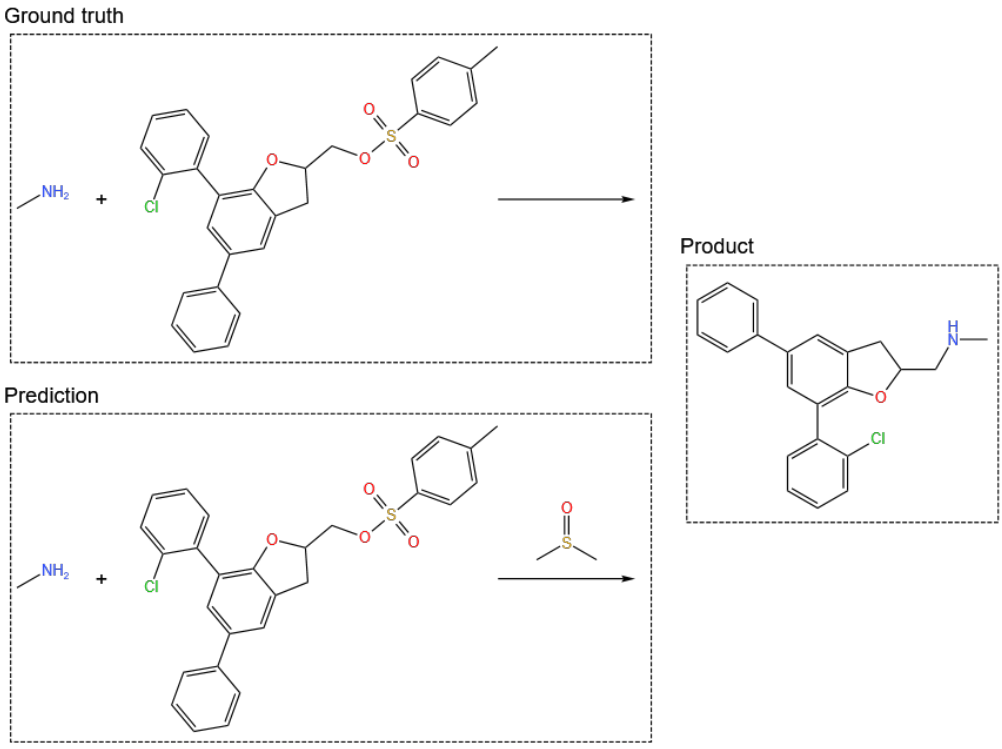

Figure 5: Case 3 from BioChem. Give the product **CNCC1Cc2cc(-c3ccccc3)cc(-c3ccccc3Cl)c2O1**, the model predicts the correct reactant **CN.Cc1ccc(S(=O)(=O)OCC2Cc3cc(-c4ccccc4)cc(-c4ccccc4Cl)c3O2)cc1**. A **catalyst S(C)(=O)C**, which bears a resemblance to the reactant structure, is predicted, possibly serving as a "reaction fragment" or an "intermediate product."

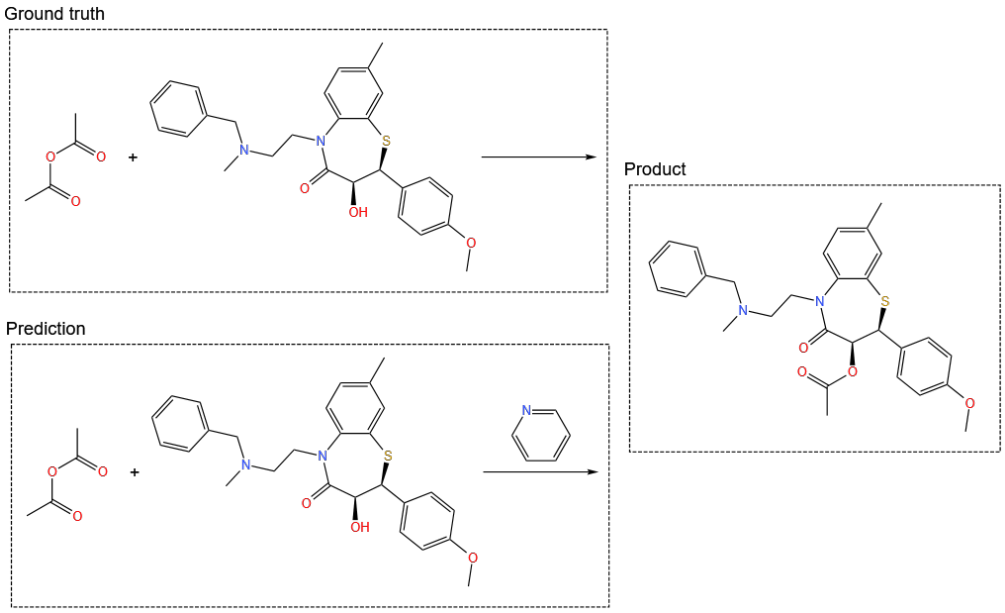

Figure 6: Case 4 from BioChem. For the product **COc1ccc([C@@H]2Sc3cc(C)ccc3N(CCN(C)Cc3ccccc3)C(=O)[C@@H]2OC(C)=O)cc1**, the model successfully predicts the reactant **CC(=O)OC(C)=O.c12ccc(C)cc1S[C@@H](c1ccc(OC)cc1)[C@@H](O)C(=O)N2CCN(C)Cc1ccccc1** and provides a catalyst **c1cccnc1**, which could be a solvent.

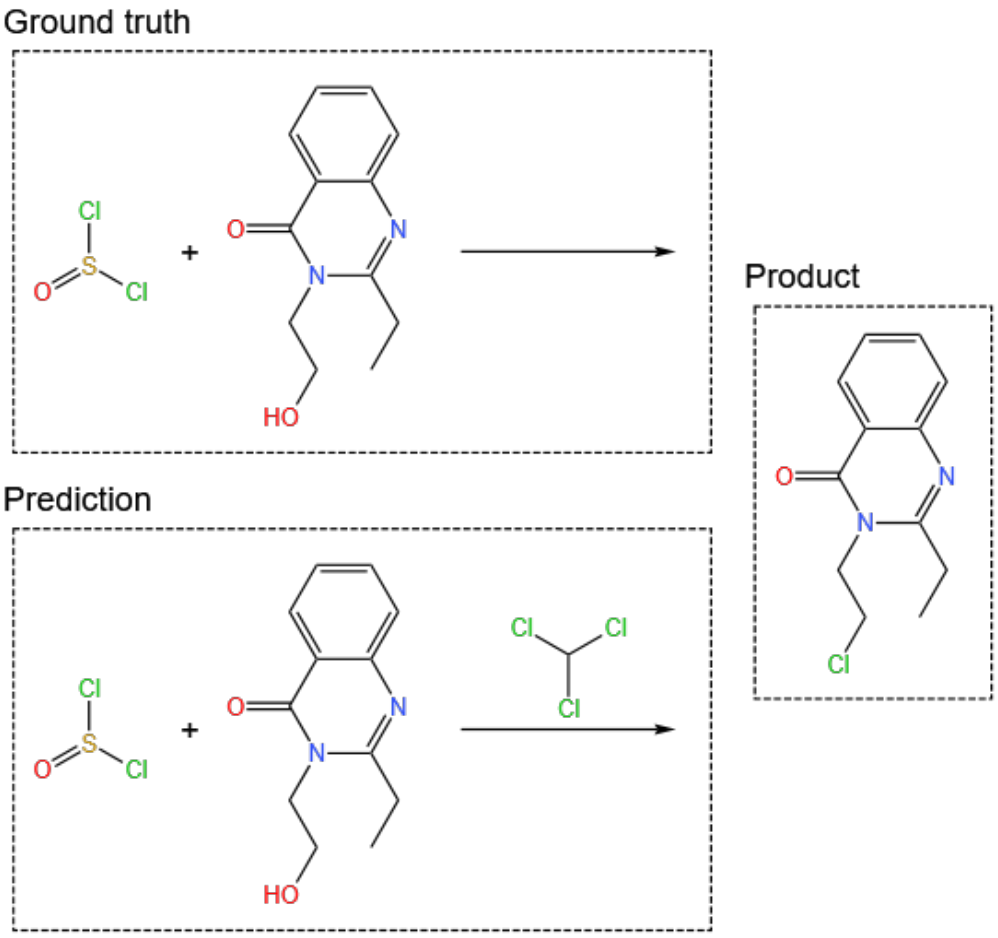

Figure 7: Case 5 from BioChem. For the product **CCc1nc2ccccc2c(=O)n1CCCl**, the model predicts the correct reactants **O=S(Cl)Cl.c12ccccc1nc(CC)n(CCO)c2=O**, and provides a catalyst **ClC(Cl)Cl**, which could be a solvent.

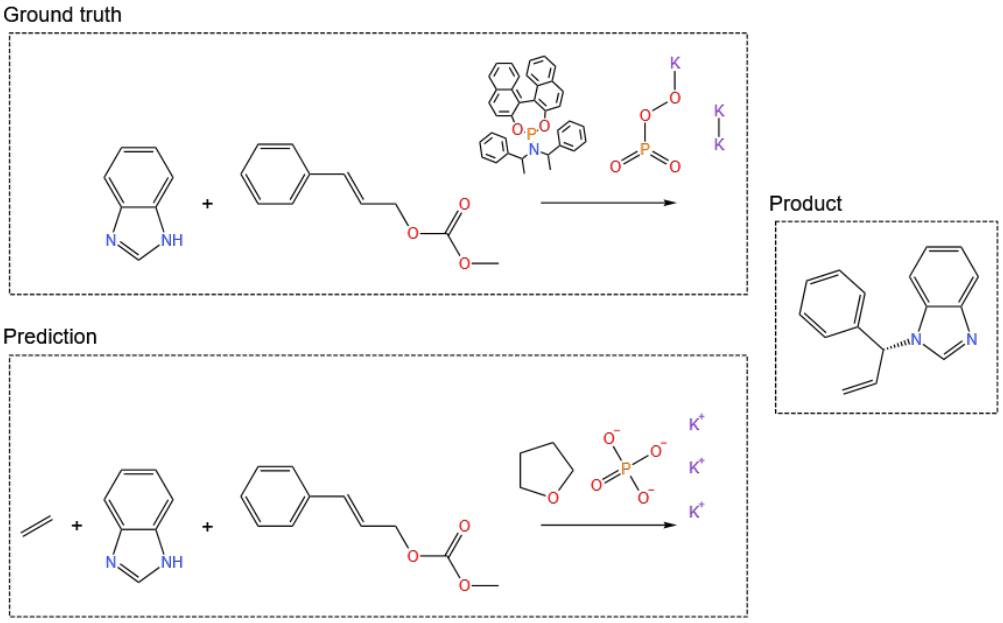

Figure 8: Case 6 from our private benchmark. For the product **C=C[C@H](c1ccccc1)n1cnc2ccccc21**, the model predicts a reactant **C=C.c12ccccc1[nH]cn2.O=C(OC/C=C/c1ccccc1)OC** with an additional small molecule ethylene, and successfully predicts furan as the solvent. The solvent information appears only in the original paper of this reaction, demonstrating that the model successfully transfers knowledge from the literature to retro-synthesis prediction tasks after being trained on a large dataset.

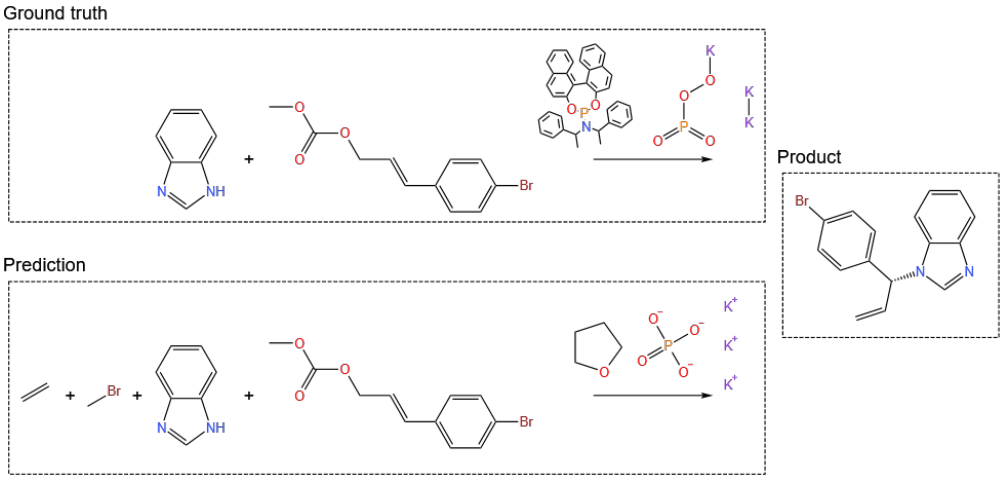

Figure 9: Case 7 from our private benchmark. For the reactant **C=C[C@H](c1ccc(Br)cc1)n1cnc2ccccc21**, the model correctly predicts the reactant **c1nc2ccccc2[nH]1.CBr.C=C.COC(=O)OC\\C=C\\c1ccc(Br)cc1** but includes an additional ethylene and its corresponding hydrogen halide molecule. The main reason might be the presence of a halogen substituent on the reactant's ring, which is relatively reasonable. The ligand is not predicted, but furan is successfully predicted as the solvent.