USPTO-LLM: A Large Language Model-Assisted Information-enriched Chemical Reaction Dataset

Shen Yuan*
Renmin University of China
Beijing, China
yuanshen721@gmail.com

Shukai Gong
Renmin University of China
Beijing, China
shukai_gong@ruc.edu.cn

Hongteng Xu†
Renmin University of China
Beijing, China
hongtengxu@ruc.edu.cn

ABSTRACT

Over the past few years, the machine learning community has given increasing attention to chemical reaction prediction and retrosynthesis. Despite impressive achievements, the existing datasets in this field have gradually become the bottleneck of current research — the limitation of dataset size and the lack of reaction condition information hinder the practicability of the current methods. In this study, we construct an information-enriched chemical reaction dataset called **USPTO-LLM**, with the help of large language models (LLMs). This dataset comprises over 247K chemical reactions extracted from the patent documents of USPTO (United States Patent and Trademark Office), encompassing abundant information on reaction conditions. We employ large language models to expedite the data collection procedures automatically with a reliable quality control process. Experiments show that USPTO-LLM helps pre-train the existing retrosynthesis methods and the condition information in the dataset helps improve the model performance. The dataset is open-sourced at https://zenodo.org/records/14396156 and the annotation code is open-sourced at https://github.com/GONGSHUKAI/USPTO_LLM.

CCS CONCEPTS

• **Applied computing** → **Chemistry**; • **Information systems** → **Data cleaning**; • **Computing methodologies** → **Spatial and physical reasoning**.

KEYWORDS

Chemical reaction data, retrosynthesis, large language model

ACM Reference Format:
Shen Yuan, Shukai Gong, and Hongteng Xu. 2018. USPTO-LLM: A Large Language Model-Assisted Information-enriched Chemical Reaction Dataset. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

*Shen Yuan and Shukai Gong have equal contributions to this work. They are listed in alphabetical order of their first names.
†The corresponding author.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The study of chemical reactions plays a central role in scientific discovery, which impacts many fields, such as chemical engineering [11], drug design [4], and material discovery [1]. The traditional research paradigm of chemical reactions heavily relies on costly and time-consuming wet lab trials and empirical expert knowledge. In the past few years, the development of artificial intelligence has provided a promising solution to accelerate the study of chemical reactions. In particular, many learning-based methods have been proposed for chemical reaction prediction [13] and retrosynthesis [5, 8], which show potential to predict chemical products and their reaction routes efficiently.

However, the learning-based methods in this field have encountered limitations and bottlenecks caused by the scarcity of high-quality data. Specifically, most existing chemical reaction datasets only focus on the reactants and products of reaction routes while ignoring the reaction conditions, e.g., catalysts, solvents, temperature, reaction duration, and so on [8]. The scarcity of information-enriched chemical reaction data leads to sub-optimal performance of learning-based methods, which prevents them from having practical applications. Take retrosynthesis (i.e., predicting chemical reaction routes based on products) as an example. The most commonly used open-source retrosynthesis dataset, USPTO-50K [12], only contains one-step chemical reactions, and only a limited number of reactions contain catalysts. As a result, the models [2, 7, 16, 21–23] trained on USPTO-50K cannot predict reactants associated with reaction conditions and thus are inapplicable in practice. In addition, given textual reaction descriptions, manually annotating complicated reaction conditions for a huge number of chemical reactions is expensive and time-consuming. Therefore, an automatic and reliable data construction method is required.

In this study, we construct a new information-enriched chemical reaction dataset called **USPTO-LLM** to overcome the above challenges. As illustrated in Figure 1, we leverage a large language model (LLM) to process the patent documents of the USPTO (the United States Patent and Trademark Office). Taking a patent document and a sophisticated prompt as input, LLM helps *i*) extract chemical reactions with reactants, products, and corresponding reaction conditions and *ii*) standardize the format of each entity. A two-round quality control process is applied to double-check the validity of the LLM's output and increase the success rate of the data construction. In this process, we verify whether the output can be formulated as a heterogeneous directed graph (HDG) and call the LLM again to process those invalid cases further, using the reasons for the invalidation as prompts. We test various LLMs to demonstrate the feasibility and universality of our data construction

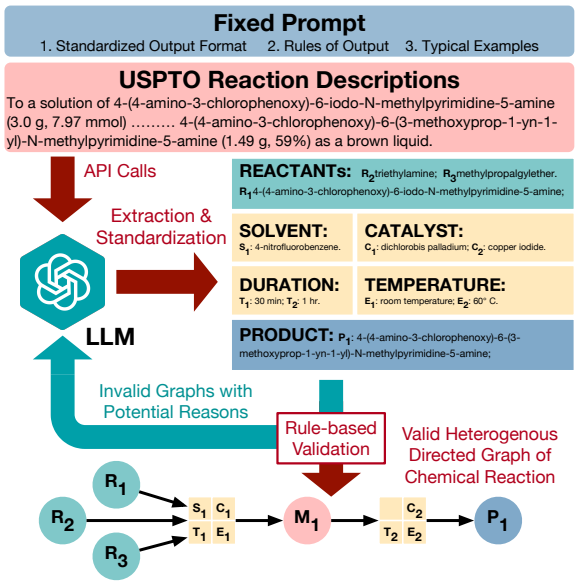

Figure 1: An illustration of our data construction method.

method. By using chemistry software such as RDKit [9] and RXN-Mapper [19], we ensure the correctness of USPTO-LLM, including the validity of chemical molecules and reaction conditions.

As a large-scale chemical reaction dataset with reaction conditions, USPTO-LLM helps develop new reaction prediction and retrosynthesis methods. Experiments demonstrate that pretraining on USPTO-LLM and incorporating reaction conditions from USPTO-LLM can improve the performance in existing retrosynthesis models. In summary, we release the dataset associated with its annotation tool, and we hope that this dataset can boost the development of artificial intelligence techniques for scientific discovery.

2 RELATED WORK

As the most commonly-used chemical reaction dataset, the original USPTO dataset [17] contains the chemical reactions extracted from United States patents published between 1976 and 2016. However, this dataset contains many duplicate reactions and erroneous information. Thus, we often apply the following three high-quality subsets in research: USPTO-50K [18] comprises 50K randomly selected reactions from USPTO and assigned reaction types and atom-to-atom mapping by the tool NameRxn [14]. USPTO-MIT, including 480K reactions, and USPTO-full, including 1M reactions, are proposed in WLDN [6] and GLN [3], respectively, after removing duplicates and erroneous reactions. However, none of the three datasets have reaction condition information. Besides the USPTO series, other chemical reaction datasets like Pistachio [15] are private and thus cannot be accessed easily.

3 THE PROPOSED USPTO-LLM DATASET

3.1 Data Structure

Given LLM's textual comprehension and re-organization capabilities, we introduce an LLM-assisted chemical reaction extraction method, transforming natural language-based chemical reaction descriptions to heterogeneous directed graphs (HDG). As illustrated in Figure 1, an HDG consists of the following two components:

- **Heterogeneous nodes:** Reactants (R), mixtures (M), and products (P) are represented as molecular nodes, while solvents (S), catalysts (C), temperatures (E), and reaction durations (T) are represented as attribute-oriented nodes.
- **Directed edges:** Directed edges linking heterogeneous nodes, with molecular nodes pointing only to attribute-oriented nodes and vice versa.

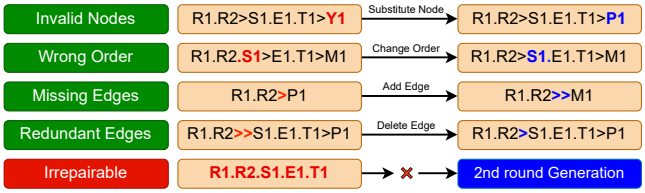

Figure 2: Illustrations of typical invalid HDGs and the rules for repairing them.

3.2 Prompt Engineering

To standardize the extraction of HDGs, besides the patent documents of chemical reactions, we apply a structured prompt as the input of LLM, which contains three parts:

- **Part 1** corresponds to the definitions of an HDG's nodes and edges and their corresponding symbol notations.
- **Part 2** contains four rules that LLM should follow when generating an HDG: (1) A standardized output format should be "$N_x.N_y{>}A{>}N$", where "." segments multiple input molecular nodes, ">" indicates directed edges, each molecular node $N \in \{R, M, P\}$, and each attribute-oriented node A is formulated as "$S.C.E.T$". (2) Each node occurs at most once in an HDG. (3) The generated HDG should reflect the correct division of reaction steps. (4) The post-processing procedures should be excluded from each step.
- **Part 3** contains five typical examples of chemical reaction extraction to support the in-context learning of LLM [10].

3.3 Two-round Generation with Quality Control

Due to the hallucination issue of LLM, some generated HDGs may violate the desired HDG structure. To balance data quality and generation cost, we use a two-round generation strategy with a feedback mechanism for quality control. In the first round, we concatenate the structured prompt with reaction descriptions and generate HDGs using LLM APIs. For those HDGs suffering invalid nodes and incorrect edges, we repair them by editing nodes or edges based on rules, as shown in Figure 2. The HDGs that can not be repaired will be regenerated by LLM, and the input of LLM combines the initial prompt, the erroneous HDGs, and the corresponding explanations for their invalidation.

The results in Table 1 verify the feasibility of the two-round generation strategy. Specifically, we apply different LLMs to generate HDGs and record the valid rate of HDG generation (i.e., the proportion of the valid HDGs in those generated by the LLMs). We find that the second-round generation helps consistently improve

LLM	GPT4 0613	GPT4 1106	GPT4 0125	GPT3.5turbo-0125
1st-round	88.1%	84.4%	85.7%	52.2%
2nd-round	51.3%	30.8%	31.8%	35.7%
Two-round	**88.3%**	**90.1%**	**90.3%**	**75.6%**

Table 1: Comparisons on the valid rate of HDGs.

the valid rate across various LLMs. According to the results, we set GPT4-0125-preview as the default LLM used in our method.

Finally, the molecules in the generated HDGs are replaced with their canonical SMILES strings and then passed through RXNMapper [19], which not only generates atom mappings for each reaction but also serves as a final quality control step. Reactions for which atom mappings cannot be created are filtered out. This process results in a total of 247K information-enriched chemical reactions.

3.4 Data Statistics

As shown in Figure 3, USPTO-LLM includes information on reaction step divisions compared to existing USPTO datasets, with 33.6% of the reactions being multi-step reactions. It also exhibits a broader distribution of reactant numbers, indicating its ability to capture a wider variety of chemical reactions compared to USPTO-50K. Figure 3 also shows the distribution of four types of reaction conditions. The top 10 catalysts and solvents in USPTO-LLM account for 81.3% and 62.4% of all catalysts and solvents. The reaction temperature and reaction duration are concentrated at room temperature (25°C) and 2×10^4 seconds respectively.

4 EXPERIMENT

4.1 Experiment Setup

We demonstrate the usefulness of USPTO-LLM in molecular retrosynthesis tasks. We consider a graph-based model **HGAR** [23] and a **Transformer**-based sequential model [20], respectively. As shown in Figure 4, HGAR is a hierarchical graph autoregressive model leveraging atom-level and motif-level information to predict reactants, while Transformer generates the SMILES strings of reactants directly given the SMILES strings of products.

To quantitatively analyze the impacts of USPTO-LLM on the above models, we designed two experiments: *i*) training the models with vs. without reaction conditions on USPTO-LLM and *ii*) training the models on USPTO-50K only vs. pretraining on USPTO-LLM and finetuning on USPTO-50K. The first experiment aims to verify the necessity of reaction condition information. In this experiment, we fuse the reaction condition information into the models by adding the embeddings of reaction conditions to the embedding of the product, i.e.,

$$\boldsymbol{M}_P \leftarrow \boldsymbol{M}_P + \mathbf{1}_N \sum\nolimits_{A \in \{S,C,E,T\}} \text{Pooling}(\boldsymbol{M}_A)^\top. \quad (1)$$

Here, $\boldsymbol{M}_P \in \mathbb{R}^{N \times d}$ represents the embedding matrix of a product P, where N is the number of graph nodes for HGAR (or the number of tokens for Transformer) and d is the dimension of embedding. For each reaction condition, i.e., $A \in \{S, C, E, T\}$, we first tokenize it at the character level into an embedding matrix $\boldsymbol{M}_A \in \mathbb{R}^{L_A \times d}$, where L_A denotes the number of characters in A, and then obtain the mean-pooling of the embedding. Each model is trained for 70 epochs with a batch size of 64. The second experiment tests the transferability of USPTO-LLM to USPTO-50K, in which each model is pretrained and finetuned for 70 epochs with a batch size of 64. For a fair comparison, the reaction types are not provided in the experiments, and we evaluate each model's performance using top-k accuracy ($k = 1, 3, 5, 10$).

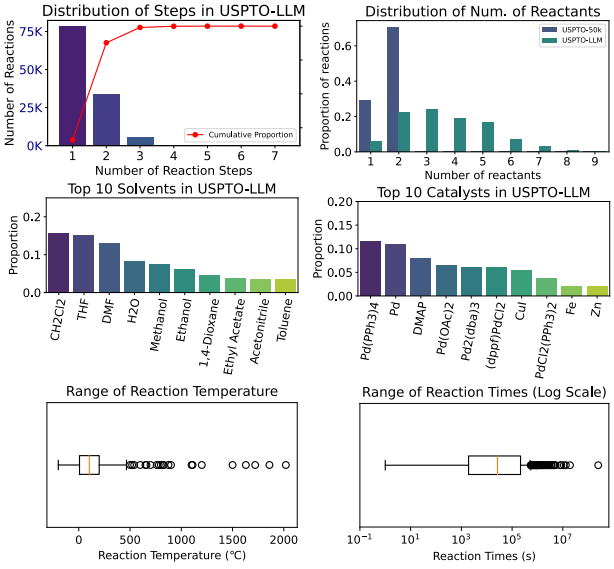

Figure 3: The distributions of reaction steps, numbers of reactants, top-10 catalysts, top-10 solvents, temperature and reaction duration.

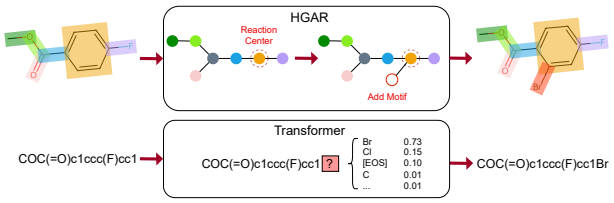

Figure 4: Illustrations of two retrosynthesis models.

4.2 Experiments and Analysis

The results in Table 2 indicate that USPTO-LLM is a challenging dataset for existing retrosynthesis models. In particular, USPTO-50K only contains simple one-step reactions, and the molecules in the dataset can be easily represented by 197 templates (i.e., typical molecular motifs or substructures). On the contrary, USPTO-LLM involves many complex multi-step reactions and contains 3,635 templates that follow a long-tailed distribution. As a result, due to the complexity of reaction types and the diversity of templates, both HGAR and Transformer suffer severe performance degradations when training and testing on USPTO-LLM. Fortunately, introducing reaction conditions helps improve the models' performance. This result demonstrates the necessity of the reaction condition information in molecular retrosynthesis tasks, which is considered in USPTO-LLM but ignored by the other datasets.

Pretraining	Training or Fine-tuning	Testing	Use Reaction Conditions	HGAR				Transformer			
				Top-1	Top-3	Top-5	Top-10	Top-1	Top-3	Top-5	Top-10
—	USPTO-LLM	USPTO-LLM	No	8.3	11.7	12.6	13.2	13.2	21.5	25.7	30.8
—	USPTO-LLM	USPTO-LLM	Yes	**10.1**	**13.0**	**13.5**	**13.9**	**15.3**	**23.5**	**27.4**	**32.2**
—	USPTO-50K	USPTO-50K	No	**55.4**	73.1	**78.6**	**83.9**	38.2	59.5	66.5	74.2
USPTO-LLM	USPTO-50K	USPTO-50K	No	53.8	**73.5**	78.0	82.7	**40.7**	**62.9**	**70.2**	**77.3**

Table 2: The performance of different models under different settings.

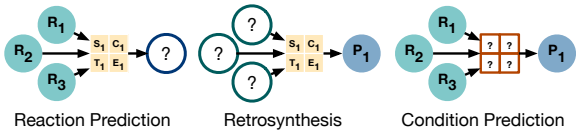

Figure 5: The HDG-based illustrations of predictive tasks.

When pretraining on USPTO-LLM, the performance of Transformer on USPTO-50K is improved across all top-k accuracy metrics, while HGAR suffers slight performance degradations on top-1, top-5, and top-10 accuracy. This phenomenon is caused by the difference between the two modeling strategies. As aforementioned, HGAR heavily relies on the leverage of molecular templates, but the distribution of the templates in USPTO-LLM is very different from that in USPTO-50K, which cannot be adapted well through the "pretraining and fine-tuning" strategy. On the contrary, Transformer models and generates SMILES strings. For USPTO-50K and USPTO-LLM, their vocabularies are overlapped by 97%, so that the sequential model pretrained on USPTO-LLM can be easily adapted to USPTO-50K and achieves encouraging performance.

5 CONCLUSION AND FUTURE WORK

We have proposed an LLM-assisted chemical reaction data extraction method and constructed an information-enriched USPTO-LLM dataset accordingly. USPTO-LLM contains rich side information such as reaction conditions and step divisions, allowing us to formulate each chemical reaction as a heterogeneous directed graph. Experiments have demonstrated that USPTO-LLM is a challenging chemical reaction dataset, which may help develop and test cutting-edge reaction prediction models.

Note that, by forming chemical reactions as HDGs, we can unify various reaction-related learning tasks in a graph-filling framework. As illustrated in Figure 5, typical reaction prediction and retrosynthesis tasks can be formulated as forward and backward node prediction tasks, respectively. Moreover, because of the availability of reaction conditions, we can even propose new learning tasks, e.g., reaction condition prediction. In the future, we plan to further enlarge the dataset and develop benchmarks for various challenging learning tasks relevant to chemical reactions based on it, including but not limited to reaction prediction, retrosynthesis, and reaction condition prediction.

REFERENCES

[1] Anthony K Cheetham, Ram Seshadri, and Fred Wudl. 2022. Chemical synthesis and materials discovery. *Nature Synthesis* 1, 7 (2022), 514–520.

[2] Ziqi Chen, Oluwatosin R Ayinde, James R Fuchs, Huan Sun, and Xia Ning. 2023. G2Retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry* 6, 1 (2023), 102.

[3] Hanjun Dai, Chengtao Li, Connor W Coley, Bo Dai, and Le Song. 2019. Retrosynthesis prediction with conditional graph logic network. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 8872–8882.

[4] Johann Gasteiger. 2007. Modeling chemical reactions for drug design. *Journal of Computer-Aided Molecular Design* 21 (2007), 33–52.

[5] Christina Humer, Rachel Nicholls, Henry Heberle, Moritz Heckmann, Michael Pühringer, Thomas Wolf, Maximilian Lübbesmeyer, Julian Heinrich, Julius Hillenbrand, Giulio Volpin, et al. 2024. CIME4R: Exploring iterative, AI-guided chemical reaction optimization campaigns in their parameter space. *Journal of Cheminformatics* 16, 1 (2024), 51.

[6] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. 2017. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems* 30 (2017).

[7] Pavel Karpov, Guillaume Godin, and Igor V Tetko. 2019. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*. Springer, 817–830.

[8] Youngchun Kwon, Dongseon Lee, Jin Woo Kim, Youn-Suk Choi, and Sun Kim. 2022. Exploring optimal reaction conditions guided by graph neural networks and Bayesian optimization. *ACS omega* 7, 49 (2022), 44939–44950.

[9] Greg Landrum et al. 2016. Rdkit: Open-source cheminformatics software, 2016. *http://www.rdkit.org* 149 (2016), 150.

[10] Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. arXiv:2212.06800 [cs.CL]

[11] Shaofen Li, Feng Xin, and Lin Li. 2017. *Reaction engineering*. Butterworth-Heinemann.

[12] Daniel Mark Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph. D. Dissertation. University of Cambridge.

[13] Ziqiao Meng, Peilin Zhao, Yang Yu, and Irwin King. 2023. A unified view of deep learning for reaction and retrosynthesis prediction: Current status and future challenges. *arXiv preprint arXiv:2306.15890* (2023).

[14] version 2.1.84 NameRxn. 2015. NextMove Software Limited. (2015). https://www.nextmovesoftware.com/namerxn.html

[15] Nextmove. 2024. Nextmove software pistachio. (2024).

[16] 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 (2021), 3273–3284.

[17] Nadine Schneider, Daniel M Lowe, Roger A Sayle, Michael A Tarselli, and Gregory A Landrum. 2016. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry* 59, 9 (2016), 4385–4402.

[18] 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 (2016), 2336–2346.

[19] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. 2021. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* 7, 15 (2021), eabe4166.

[20] 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 (2020), 5575.

[21] Yue Wan, Chang-Yu Hsieh, Ben Liao, and Shengyu Zhang. 2022. Retroformer: Pushing the Limits of End-to-end Retrosynthesis Transformer. In *International Conference on Machine Learning*. PMLR, 22475–22490.

[22] Xiaorui Wang, Yuquan Li, Jiezhong Qiu, Guangyong Chen, Huanxiang Liu, Benben Liao, Chang-Yu Hsieh, and Xiaojun Yao. 2021. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal* 420 (2021), 129845.

[23] Shen Yuan, Fanmeng Wang, Zhewei Wei, Peilin Zhao, Lanqing Li, and Hongteng Xu. 2024. Learning A Hierarchical Graph Autoregression Model for Semi-template Molecular Retrosynthesis. (2024).