ReactIE: Enhancing Chemical Reaction Extraction with Weak Supervision

Ming Zhong Siru Ouyang Minhao Jiang
Vivian Hu Yizhu Jiao Xuan Wang Jiawei Han
University of Illinois Urbana-Champaign, IL, USA
{mingz5, siruo2, minhaoj2, vivianhu2, yizhuj2, xwang174, hanj}@illinois.edu

Abstract

Structured chemical reaction information plays a vital role for chemists engaged in laboratory work and advanced endeavors such as computer-aided drug design. Despite the importance of extracting structured reactions from scientific literature, data annotation for this purpose is cost-prohibitive due to the significant labor required from domain experts. Consequently, the scarcity of sufficient training data poses an obstacle to the progress of related models in this domain. In this paper, we propose REACTIE, which combines two weakly supervised approaches for pre-training. Our method utilizes frequent patterns within the text as *linguistic cues* to identify specific characteristics of chemical reactions. Additionally, we adopt synthetic data from patent records as distant supervision to incorporate *domain knowledge* into the model. Experiments demonstrate that REACTIE achieves substantial improvements and outperforms all existing baselines.

1 Introduction

The integration of advanced Natural Language Processing (NLP) techniques in the field of chemistry has been gaining significant attention in both academia and industry (Wang et al., 2019; Fabian et al., 2020; Chithrananda et al., 2020). By formulating applications in chemistry as molecular representation (Shin et al., 2019; Wang et al., 2022a), information extraction (Vaucher et al., 2020; Wang et al., 2021, 2022b), and text generation (Edwards et al., 2022) tasks, NLP approaches provide new avenues for effective understanding and analysis of chemical information. In particular, we focus on the chemical reaction extraction task, as it can serve as a valuable reference for chemists to conduct bench experiments (Guo et al., 2022).

Despite the abundance of text describing chemical reactions in the scientific literature, the conversion to a structured format remains a major challenge. One approach is the utilization of domain

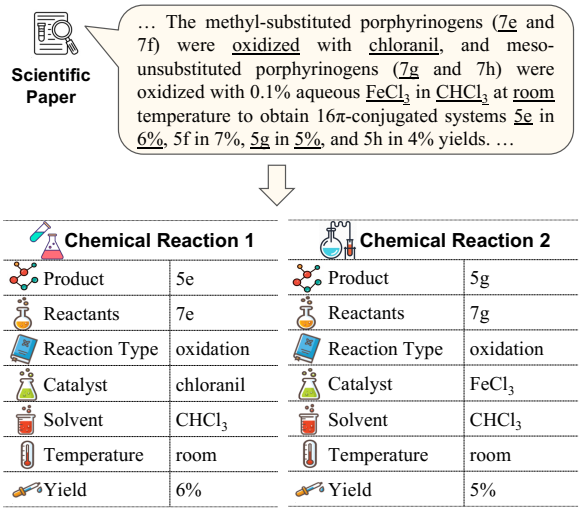

Figure 1: An example of the chemical reaction extraction task. This figure depicts two out of the four chemical reactions present in the text for simplicity. The passage is drawn from Ahmad et al. (2015).

experts to manually extract chemical reactions, resulting in several commercial reaction databases, such as Reaxys (Goodman, 2009) and SciFinder (Gabrielson, 2018). However, this method is associated with significant time and labor costs, as well as the issue of restricted access to these resources.

Subsequently, research efforts concentrated on automated systems, including OPSIN (Lowe, 2012) and CHEMRXNBERT (Guo et al., 2022). OPSIN is a heuristic-based system that employs a complex set of rules to identify the reaction roles. While it is effective for well-formatted text, OPSIN's performance is limited in scientific literature due to its sensitivity to variations in language use. In contrast, Guo et al. (2022) obtained CHEMRXNBERT by pre-training with language modeling on chemistry journals, however, the model performance is constrained by the small size of the training set during fine-tuning. This raises the question of how to effectively utilize large-scale unlabeled data for this task, which remains an under-explored area.

In this paper, we present REACTIE, a pre-trained

model for chemical reaction extraction. In light of the clear gap between prevalent pre-training tasks and the applications in the field of chemistry, we propose two weakly supervised methods to construct synthetic data for pre-training. Intuitively, humans can infer certain roles in chemical reactions from **linguistic cues**. As shown in Figure 1, we can identify "*5e*" as the product from the semantic meaning of the phrase "*to obtain 5e*". To this end, we mine frequent patterns from texts as linguistic cues and inject them into the model. Furthermore, **domain knowledge** also plays a crucial role in this task. For example, the accurate identification of "*chloranil*" as a catalyst rather than a reactant in Figure 1 requires a deep understanding of related compounds. To address this, we incorporate domain knowledge into REACTIE by utilizing patent literature as distant supervision. By pre-training on these acquired synthetic data, REACTIE maintains consistency with downstream objectives.

Experimentally, REACTIE achieves state-of-the-art performance, improving F_1 scores by 14.9 and 2.9 on the two subtasks, respectively. Moreover, we conduct ablation studies to examine the contributions of the proposed methods. Fine-grained analyses are performed to investigate the effects of pre-training strategies on different reaction roles. Our findings suggest that linguistic cues are crucial for extracting products and numbers, while chemical knowledge plays an essential role in understanding catalysts, reactants, and reaction types.

2 Preliminary

2.1 Task Formulation

Given a text D, the goal of this task is to extract all the structured chemical reactions $\mathcal{S}$ in D, where each $S \in \mathcal{S}$ contains n role-argument pairs $\{(r_1, a_1), \cdots, (r_n, a_n)\}$. The roles are 8 pre-defined attributes in a chemical reaction, including *product*, *reactant*, *catalyst*, *solvent*, *reaction type*, *temperature*, and *yield*. Each S does not include the roles that are not present in the original text. Definitions for each role are included in Appendix A.

2.2 Workflow for IE System

From the perspective of the model, existing systems typically follow a two-step pipeline:

1) **Product Extraction**: In chemical reactions, the product is the central factor as the same reactants can yield varying products depending on the reaction conditions. Therefore, the IE systems first extract all the products in D to determine the number of chemical reactions, i.e., the number of S. This step can also be used to extract passages in a scientific paper that contain chemical reactions.

2) **Role Extraction**: Given the original text D and the specific product, the IE systems are required to capture the relationship between the entities in D and the product, extract the corresponding reaction roles, and output the final $\mathcal{S}$.

3 REACTIE Framework

3.1 Reformulation

Previous studies have defined this task as a sequence labeling problem[1]. However, this approach could be inadequate in certain cases. For instance, the final argument may be an alias, abbreviation, or pronoun of a compound in D, or the necessary conversion of words should be made (as illustrated in Figure 1, "*oxidized*" $\rightarrow$ "*oxidation*").

In light of these limitations, we reformulate the chemical reaction extraction task as a Question Answering (QA) problem, utilizing the pre-trained generation model FLAN-T5 (Chung et al., 2022) as the backbone. For product extraction, the input question is "*What are the products of the chemical reactions in the text?*". For role extraction, such as catalyst, the corresponding question is "*If the final product is X, what is the catalyst for this chemical reaction?*". In this unified QA format, we present the pre-training stage of REACTIE as follows.

3.2 Pre-training for REACTIE

Given the clear discrepancy between prevalent pre-training tasks such as language modeling and the task of chemical reaction extraction, we propose two weakly supervised methods for constructing synthetic data to bridge this gap.

Linguistics-aware Data Construction Intuitively, it is possible for humans to infer certain properties of a chemical reaction, even without any prior knowledge of chemistry. As an example, consider the sentence "*Treatment of 13 with lithium benzyl oxide in THF afforded the dihydroxybenzyl ester 15*" (Dushin and Danishefsky, 1992). We can identify that "*13*" and "*lithium benzyl*" are the reactants, and "*dihydroxybenzyl ester 15*" is the end product, without knowing any specific compounds involved. This can be achieved by utilizing **linguis-**

[1]The reaction roles are captured using "BIO" scheme.

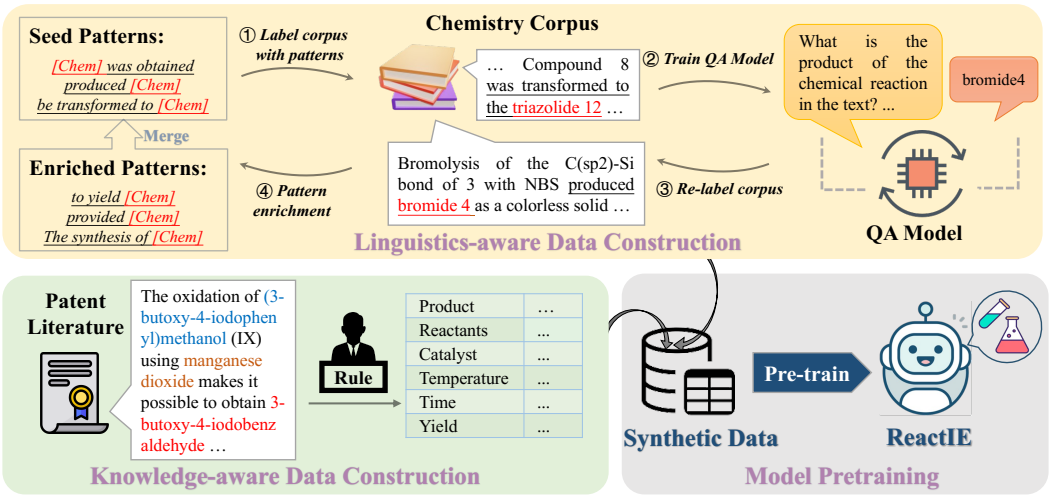

Figure 2: Overview of REACTIE. We propose linguistics-aware and knowledge-aware methods to construct synthetic data, thus bridging the gap between the objectives of pre-training and the chemical reaction extraction task.

tic cues such as the semantics of phrases and the structure of sentences to extract the arguments.

Inspired by this, we leverage frequent patterns (Jiang et al., 2017) in the text that describes specific reaction roles as linguistic cues. Take product extraction as an example, we first replace the chemical with a special token "[Chem]" using CHEMDATAEXTRACTOR (Swain and Cole, 2016), and then manually create a set of seed patterns, such as *the produced [Chem]*, *conversion of [Chem] to [Chem]*, etc. The red [Chem] indicates that the chemical here is the product of a reaction. As shown in Figure 2, based on seed patterns and a chemistry corpus, we construct synthetic data as:
1) Seed patterns are used to annotate the chemical corpus, resulting in training data containing labels.
2) Continue training Flan-T5 in QA format on the data from the previous step.
3) Use the QA model to re-label the entire corpus.
4) The most frequent patterns are mined from the data in step 3 as the enriched pattern set.

By merging the seed patterns in the first step with the enriched patterns, we can iteratively repeat the process and collect reliable data containing multiple linguistic cues. More examples and details can be found in Appendix B and Table 4.

Knowledge-aware Data Construction In addition to utilizing linguistic cues, a deep understanding of chemical reactions and terminology is imperative for accurately extracting information from texts. This is exemplified in the case presented in Figure 1, in which the roles of compounds such as "*chloranil*", "$FeCl_3$" and "$CHCl_3$" as reactants, catalysts, or solvents cannot be inferred without prior knowledge. In light of this, we propose the integration of **domain knowledge** into REACTIE through the synthetic data derived from patent records.

The text within patent documents is typically well-formatted, allowing for the extraction of structured chemical reactions through the well-designed rules incorporating multiple chemical principles and associated knowledge bases (Lowe, 2012). To utilize this, we adopt datasets extracted from the U.S. patent literature by OPSIN (Lowe, 2018) as our synthetic data. We focus on 4 reaction roles (product, reactant, catalyst, and solvent) that are most relevant to chemistry knowledge.

Training Paradigm The methods outlined above enable the acquisition of a substantial amount of synthetic data. We then proceed to conduct pre-training by building upon the FLAN-T5 model in a text-to-text format. The input contains questions q_i specific to a reaction role r_i and text D, and the output is the corresponding argument a_i or "None". After pre-training, the unsupervised version of REACTIE acquires the capability to extract structured chemical reactions. To further improve it, we also perform fine-tuning on an annotated dataset to attain a supervised version of REACTIE.

4 Experiments

4.1 Experimental Setup

Datasets We use Reaction Corpus (Guo et al., 2022) which includes 599/96/111 annotated chemical reactions in training, dev, and test sets. The input is a paragraph in scientific papers and the output consists of multiple structured chemical reactions

Models	P (%)	R (%)	F (%)
Unsupervised			
OPSIN	18.8	5.4	8.4
REACTIE	**69.7**	**53.5**	**60.5**
Supervised			
BILSTM	52.4	46.7	49.4
BILSTM (w/ CRF)	54.3	49.1	51.6
BERT	78.8	56.8	66.0
BIOBERT	76.4	61.3	68.0
CHEMBERT	84.6	69.4	76.2
FLANT5	88.0	83.2	85.5
REACTIE	**94.2**	**88.2**	**91.1**
- *linguistics cues*	89.8	84.7	87.2
- *domain knowledge*	92.6	87.1	89.8

Table 1: Results for product extraction. The results presented in the gray background correspond to the performance of REACTIE and its ablation studies.

in the text. This corpus is designed to evaluate two subtasks, product extraction, and role extraction.

Baselines We compare the performance of REACTIE with several state-of-the-art baselines, including OPSIN, BILSTM-CRF (Huang et al., 2015), BERT (Devlin et al., 2019), BIOBERT (Lee et al., 2020), CHEMBERT, and CHEMRXNBERT (Guo et al., 2022). OPSIN is an unsupervised rule-based system while the variants of BERT are pre-trained on different domain-specific corpora.

Implementation Details We use "google/flan-t5-large" as the backbone model in all experiments. For linguistics-aware data construction, we perform 3 iterations on 18,894 chemical journals and end up with 92,371 paragraphs containing the linguistic cues of product, temperature, yield, and time. Other reaction roles are excluded because they do not have sufficient patterns to ensure the reliability of the data. For knowledge-aware data construction, excessively long (> 256 words) and short (< 8 words) texts, as well as samples where the arguments do not appear in the original text, are filtered to yield 100,000 data. We train REACTIE for 1 epoch with 0.1 label smoothing on a total of 192,371 samples. For both pre-training and fine-tuning, we set the batch size to 16 with 5e-5 as the learning rate. All results are the performance of the checkpoints selected by the dev set.

4.2 Experimental Results

Results for Product Extraction The first part of Table 1 presents the results under the unsupervised setting. OPSIN performs poorly in the scientific paper domain due to its sensitivity to language usage. In contrast, REACTIE demonstrates superior extraction capabilities after pre-training and outperforms the fully supervised BiLSTM (w/ CRF).

Under the supervised setting, REACTIE attains state-of-the-art performance with a significant margin, achieving a 14.9 increase in F_1 scores compared to CHEMBERT. While our backbone model, FLANT5, shows outstanding results, our proposed methods can lead to further gains (85.5 $\Rightarrow$ 91.1 F_1). Ablation studies highlight the importance of linguistics-aware pre-training over in-domain knowledge in the product extraction subtask. This finding also supports the advantages of pre-trained language models (FLANT5) over domain-specific models (CHEMBERT), as the writers have provided sufficient linguistic cues for the products of chemical reactions when describing them.

Models	P (%)	R (%)	F (%)
BERT	69.2	69.2	69.2
BIOBERT	73.3	75.5	74.3
CHEMBERT	77.0	76.4	76.7
CHEMRXNBERT	79.3	78.1	78.7
FLANT5	76.1	75.4	75.8
REACTIE	**80.8**	**82.5**	**81.6**
- *linguistics cues*	78.1	83.3	80.6
- *domain knowledge*	74.8	79.8	77.2

Table 2: Results for role extraction.

Results for Role Extraction As listed in Table 2, REACTIE also beats the previous best model CHEMRXNBERT by 2.9 F_1 score for the role extraction subtask. In comparison to the product, the accurate extraction of other reaction roles from the original text necessitates a greater level of in-domain knowledge. Specifically, the model performance decreases slightly (81.6 $\Rightarrow$ 80.6 F_1) when linguistics-aware pre-training is removed, and substantially by 4.4 (81.6 $\Rightarrow$ 77.2 F_1) when knowledge-aware pre-training is no longer incorporated. The results of these two subtasks reveal that our proposed approaches are complementary and indispensable in enabling REACTIE to fully comprehend chemical reactions. Together, they contribute to a deeper understanding of the task from both linguistic and chemical knowledge perspectives.

Analysis for Reaction Roles To further investigate the effect of our pre-training strategies, we present ΔF_1 scores on different reaction roles after equipping the two methods separately in Figure 3. We can observe that these two strategies assist

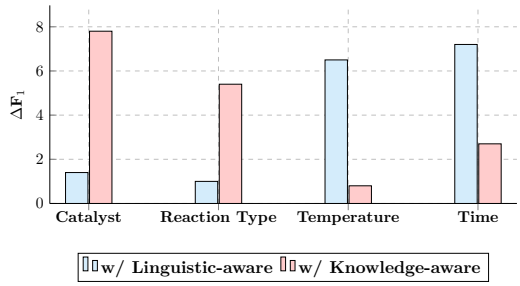

Figure 3: The impact of two pre-training strategies on different chemical reaction roles. The Y-axis shows the F_1 improvement compared to the backbone model.

the model by concentrating on distinct aspects of chemical reactions. Linguistic-aware pre-training primarily improves performance in reaction roles related to numbers, as these numbers tend to appear in fixed meta-patterns. In contrast, knowledge-related pre-training significantly enhances the results of catalyst and reaction type, which require a chemical background for accurate identification. Overall, the combination of both approaches contributes to the exceptional performance of REACTIE in the chemical reaction extraction task.

5 Conclusion

In this paper, we present REACTIE, an automatic framework for extracting chemical reactions from the scientific literature. Our approach incorporates linguistic and chemical knowledge into the pre-training. Experiments show that REACTIE achieves state-of-the-art results by a large margin.

Limitations

We state the limitations of this paper from the following three aspects:

1) Regarding linguistics-aware data construction, we only perform seed-guided pattern enrichment for four reaction roles (product, yield, temperature, and time, see Table 4) due to the lack of sufficient reliable patterns for other roles. Incorporating more advanced pattern mining methods (Li et al., 2018; Chen et al., 2022) may alleviate this issue and discover more reliable linguistic cues, which we leave for future work.

2) As in the previous work, we adopt a fixed reaction scheme to extract structured chemical reaction information. However, there are always new informative roles in the text (Jiao et al., 2022), such as experimental procedures (Vaucher et al., 2021), so how to predict both roles and arguments without being limited to a fixed scheme could be a meaningful research topic.

3) REACTIE is capable of detecting chemical reactions within scientific literature by predicting if a given passage contains a product. However, accurate text segmentation of a paper remains an unresolved and crucial issue. Incomplete segmentation may result in the failure to fully extract reaction roles, while excessively long segmentation may negatively impact the model performance. Therefore, integrating a text segmentation module into the existing two-step pipeline may be the next stage in the chemical reaction extraction task.

Acknowledgements

We thank anonymous reviewers for their valuable comments and suggestions. Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

Sohail Ahmad, Kumar Karitkey Yadav, Soumee Bhattacharya, Prashant Chauhan, and SMS Chauhan. 2015. Synthesis of 21, 23-selenium- and tellurium-substituted 5-porphomethenes, 5, 10-porphodimethenes, 5, 15-porphodimethenes, and porphotrimethenes and their interactions with mercury. *The Journal of Organic Chemistry*, 80(8):3880–3890.

Yulong Chen, Yang Liu, Li Dong, Shuohang Wang, Chenguang Zhu, Michael Zeng, and Yue Zhang. 2022. Adaprompt: Adaptive model training for prompt-based NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 6057–6068. Association for Computational Linguistics.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Russell G Dushin and Samuel J Danishefsky. 1992. Total syntheses of ks-501, ks-502, and their enantiomers. *Journal of the American Chemical Society*, 114(2):655–659.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, and Heng Ji. 2022. Translation between molecules and natural language. *CoRR*, abs/2204.11817.

Benedek Fabian, Thomas Edlich, Héléna Gaspar, Marwin Segler, Joshua Meyers, Marco Fiscato, and Mohamed Ahmed. 2020. Molecular representation learning with language models and domain-relevant auxiliary tasks. *arXiv preprint arXiv:2011.13230*.

Stephen Walter Gabrielson. 2018. Scifinder. *Journal of the Medical Library Association: JMLA*, 106(4):588.

Jonathan M. Goodman. 2009. Computer software review: Reaxys. *J. Chem. Inf. Model.*, 49(12):2897–2898.

Jiang Guo, A. Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W. Coley, Klavs F. Jensen, and Regina Barzilay. 2022. Automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model.*, 62(9):2035–2045.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M. Kaplan, Timothy P. Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 877–886. ACM.

Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. 2022. Open-vocabulary argument role prediction for event extraction. *CoRR*, abs/2211.01577.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Qi Li, Meng Jiang, Xikun Zhang, Meng Qu, Timothy P. Hanratty, Jing Gao, and Jiawei Han. 2018. Truepie: Discovering reliable patterns in pattern-based information extraction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 1675–1684. ACM.

Daniel Lowe. 2018. Chemical reactions from us patents (1976-sep2016). *doi*, 10:m9.

Daniel M. Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge, UK.

Bonggun Shin, Sungsoo Park, Keunsoo Kang, and Joyce C. Ho. 2019. Self-attention based molecule representation for predicting drug-target interaction. In *Proceedings of the Machine Learning for Healthcare Conference, MLHC 2019, 9-10 August 2019, Ann Arbor, Michigan, USA*, volume 106 of *Proceedings of Machine Learning Research*, pages 230–248. PMLR.

Matthew C. Swain and Jacqueline M. Cole. 2016. Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature. *J. Chem. Inf. Model.*, 56(10):1894–1904.

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):1–11.

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):1–11.

Hongwei Wang, Weijiang Li, Xiaomeng Jin, Kyunghyun Cho, Heng Ji, Jiawei Han, and Martin D. Burke. 2022a. Chemical-reaction-aware molecule representation learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, BCB 2019,*

Niagara Falls, NY, USA, September 7-10, 2019, pages 429–436. ACM.

Xuan Wang, Vivian Hu, Minhao Jiang, Yu Zhang, Jinfeng Xiao, Danielle Cherrice Loving, Heng Ji, Martin Burke, and Jiawei Han. 2022b. REACTCLASS: cross-modal supervision for subword-guided reactant entity classification. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2022, Las Vegas, NV, USA, December 6-8, 2022*, pages 844–847. IEEE.

Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. Chemner: Fine-grained chemistry named entity recognition with ontology-guided distant supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5227–5240. Association for Computational Linguistics.

Reaction Role	Description
Product	Chemical substance that is the final outcome (major product) of the reaction
Reactants	Chemical substances that contribute heavy atoms to the product
Catalyst	Chemical substances that participate in the reaction but do not contribute heavy atoms (e.g., acid, base, metal complexes)
Solvent	Chemical substances that are used to dissolve/mix other chemicals, typically quantified by volume and used in superstoichiometric amounts (e.g., water, toluene, THF)
Temperature	Temperature at which the reaction occurs
Time	Duration of the reaction performed
Reaction Type	Descriptions about the type of chemical reaction
Yield	Yield of the product

Table 3: Reaction scheme used in this paper.

A Reaction Scheme

We adopt the same reaction scheme as in the previous study, including 8 pre-defined reaction roles to cover the source chemicals, the outcome, and the conditions of a chemical reaction. To help better understand each reaction role, we include the detailed descriptions of the reaction scheme in Guo et al. (2022) as a reference in Table 3.

B Pattern Enrichment in Linguistics-aware Data Construction

Table 4 provides examples of seed and enriched patterns for the product, yield, temperature, and time. In each iteration, we extract n-grams ($n = \{2, \cdots, 6\}$) containing the product ([Chem]), yield ([Num]), temperature ([Num]), and time ([Num]) from the corpus re-labeled by the QA model and remove the redundant patterns. We manually review and select reliable patterns and merge them into the pattern set of the previous iteration.

Seed Patterns (completed set)	**Enriched Patterns** (randomly sampled set)
Product	
produced [Chem]	to yield [Chem]
[Chem] be obtained	provided [Chem]
[Chem] be transformed to [Chem]	synthesis of [Chem]
[Chem] be systhesized from [Chem]	[Chem] be prepared from [Chem]
conversion of [Chem] to [Chem]	desired [Chem]
Yield	
in [Num] % yield	at [Num] % conversion
a yield of [Num] %	in [Num] % isolated yield
([Num] % yield)	([Num] % overall)
Temperature	
at [Num] °C	([Num] °C)
at [Num] K	a reaction temperature of [Num] °C
at [Num] OC	from [Num] to [Num] °C
Time	
for [Num] h	over [Num] h
for [Num] min	within [Num] h
for [Num] seconds	([Num] °C, [Num] h)
after [Num] h	for [Num] days

Table 4: Examples of Seed and enriched meta-patterns for the product, yield, and temperature. All seed patterns (3-5 patterns per role) are included here, and random samples from the enriched patterns are used as examples (20-50 patterns total per role). The red [Chem] indicates that the corresponding chemical is a product. The blue [Num] denotes that the number is a yield, green [Num] represents the temperature, and yellow [Num] is the reaction time.