Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Digital Discovery*, 2025, **4**, 172

A framework for reviewing the results of automated conversion of structured organic synthesis procedures from the literature†

Kojiro Machi, *[a] Seiji Akiyama, [b] Yuuya Nagata [b] and Masaharu Yoshioka *[abc]

Organic synthesis procedures in the scientific literature are typically shared in prose (*i.e.*, as unstructured data), which is not suitable for data-driven research applications. To represent such procedures, there is a well-structured language, named chemical description language (χDL). While automated conversion methods from text to χDL using either a rule-based approach or a generative large language model (GLLM) have been proposed, they sometimes produce errors. Therefore, human review following an automated conversion is essential to obtain an accurate χDL. The aim of this work is to visualize embedded information in the original text with a structured format to support the understanding of human reviewers. In this paper, we propose a novel framework for editing automatically converted χDLs from the literature with annotated text. In addition, we introduce a rule-based conversion method. To improve the quality of automated conversions, a method of using two candidate χDLs with different characteristics was proposed: one generated by the proposed rule-based method and the other by an existing GLLM-based method. In an experiment involving six organic synthesis procedures, we confirmed that showing the outputs of both systems to the user improved recall compared with showing one output individually.

Received 18th October 2024
Accepted 25th November 2024

DOI: 10.1039/d4dd00335g

rsc.li/digitaldiscovery

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

1 Introduction

Organic synthesis procedures in the scientific literature are typically shared in prose (*i.e.*, as unstructured data). Reproducing these procedures is sometimes challenging because these texts can be ambiguous and require interpretation by human experts. While recent machine learning techniques and laboratory automation can accelerate chemical research,[1] the absence of machine-readable procedures hinders the application of these efforts. If these procedures are shared as an findable, accessible, interoperable, and reusable (FAIR) format,[2] it will help not only computers but also people who want to execute experiments but are not familiar with organic synthesis.

There are several schemes for representing organic synthesis procedures and these can be classified into two levels: (a) a general description that cannot be executable on the platforms[3–8] and (b) a detailed description that can be executable on robotic platforms.[9,10] At the detailed level, Mehr *et al.* proposed the chemical description language (χDL).[10] χDL was designed as a universal chemical programming language that could be executed on any automated platform by translating into platform-specific low-level actions if the actions were feasible on the platforms. Their research group demonstrated the capability of χDL by executing chemical reactions on their robotic platform. Furthermore, several examples that use χDL to execute automated chemical reactions on other platforms have been reported.[11,12] An integrated development environment for χDL named ChemIDE and a rule-based natural language processing (NLP) tool for the conversion of organic synthesis procedures from text to χDL have also been proposed.

Because manual information extraction from the chemical literature is a labor-intensive task for domain experts, NLP tools have been developed to support this work. From an early stage, rule-based methods have been developed for the extraction of information, such as compound names, reaction parameters, and actions.[3,13–15] For example, ChemicalTagger[3] was developed to extract chemical reaction information from the literature. Because the extraction is done by rules, domain experts can see the alignment of raw text and the output if it is visualized. However, rule-based methods sometimes suffer from scalability and flexibility issues against a wide variety of texts. In the past decade, deep learning-based methods have also been developed.[16,17] These methods are generally more flexible and robust to data variations and show higher performance than rule-based methods but require large amounts of task-specific data for the training of a model. To enable the training of deep learning models with smaller datasets, bidirectional encoder

[a]*Graduate School of Information Science and Technology, Hokkaido University, Kita 14 Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan*
[b]*Institute for Chemical Reaction Design and Discovery (WPI-ICReDD), Hokkaido University, Kita 21 Nishi 10, Kita-ku, Sapporo, Hokkaido 001-0021, Japan*
[c]*Faculty of Information Science and Technology, Hokkaido University, Kita 14 Nishi 9, Kita-ku, Sapporo, Hokkaido, 060-0814, Japan. E-mail: yoshioka@ist.hokudai.ac.jp*
† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00335g

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

representations from transformers (BERT)[18] were proposed. BERT employed self-supervised pretraining to obtain general natural language patterns and supervised fine-tuning for solving a specific task. To obtain higher performance in domain-specific text, several domain-specific BERTs have been introduced.[19–23] For chemistry, ChemBERT[22] was proposed. In the past few years, generative large language models (GLLMs), which have shown high performance when trained only with zero or a few training examples, have constituted a trend in NLP tasks.[24–26] For chemical information extraction, several works demonstrated the usefulness of GLLMs.[11,27–29] However, several challenges, including consistency of the output format and unclear text alignment, hinder the wide application of these models.

For the automated conversion from text to χDL, Yoshikawa *et al.* proposed a GLLM-based method named CLAIRify[11] and compared the performance of CLAIRify with a rule-based SynthReader. CLAIRify employed an iteration cycle of generation by a GLLM (GPT3.5, one of the GPT models[31]) and the validation of generated code to obtain a syntactically correct output. The outputs of organic synthesis procedures generated by SynthReader and CLAIRify were compared by expert chemists, and the experts often preferred the outputs of CLAIRify over those from SynthReader. In addition, CLAIRify tends to obtain higher recall and lower precision than SynthReader. The experts mentioned that the effects of missing actions are more severe than ambiguous or wrong actions when they determine preferred outputs.

While the focus of these studies was on improving the performance of automatic extraction, the aspect of manually correcting the automatically extracted results was not systematically investigated. However, a human review process is essential to make appropriate χDLs from the literature because these methods did not have 100% accuracy. In the review step, human reviewers need to read the original text.

Along these lines, the aim of this work is to visualize embedded information in the original text with a structured format to support the understanding of human reviewers. We propose a novel framework for editing automatically converted χDLs from the literature with annotated text. Fig. 1 shows the overview of our framework. Our framework has two main points. First, to make actions described in plain text easier to understand visually, our framework provides reviewers with annotated text; it annotates action verbs with related entities and parameters. We used the organic synthesis procedures with argument roles (OSPAR) format,[8] which was developed in our previous work, as the annotation format. Here, the structuring of procedures is aimed only at the synthesis sections, and no structuring is performed for the purification or analysis sections reported in the literature. This is because purification and analysis involve a wide variety of operations and require the description of equipment-specific procedures, which are currently considered unsuitable for structuring. Consequently, the conversion of text to χDL is also restricted to the synthesis sections in this study. Second, to improve automated conversion quality, we propose a method to use two candidate χDLs with different characteristics. One is the GLLM-based CLAIRify and the other is a rule-based system that was developed in this study. Although CLAIRify achieved higher recall than the rule-based system, there are several cases where only the rule-based systems can find appropriate information. Therefore, it is useful for the user to refer to both results to select appropriate

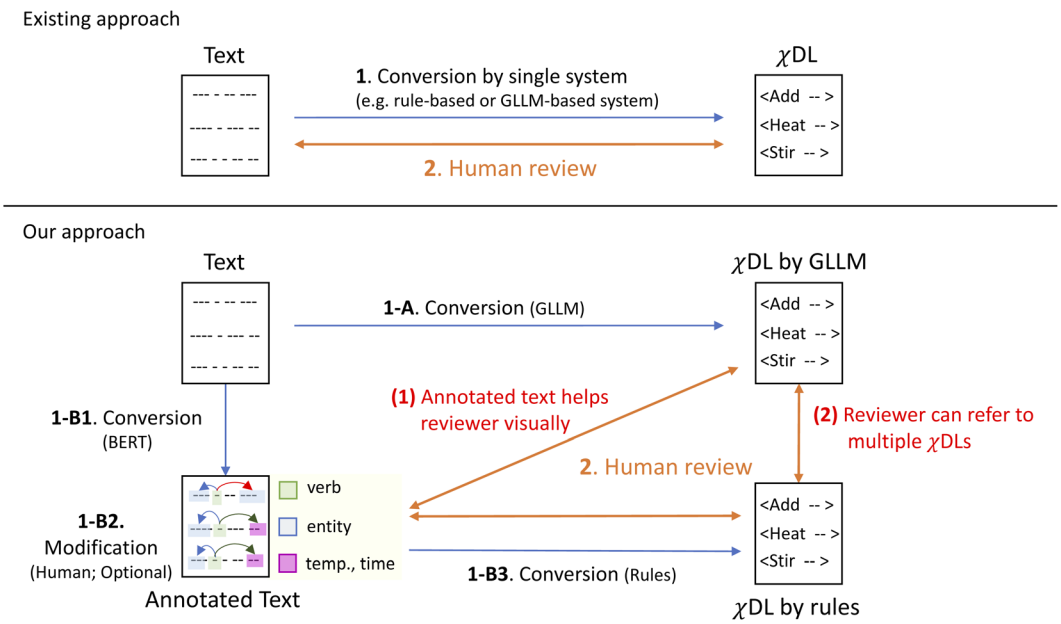


Fig. 1 Overview of our approach. In existing approaches, χDL is generated by a single system, followed by a human review that compares it with unannotated text. In contrast, our approach generates χDL using two systems: (A) a GLLM-based system and (B) a rule-based proposed system. In the GLLM-based system, χDL is directly transformed from the text (1-A). In the proposed system, text annotation is conducted using BERT (1-B1), and if necessary, the annotations can be modified (1-B2). Then, the annotated text is converted into χDL by a rule-based method (1-B3). Finally, human reviewers can compare the converted χDLs with annotated text.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

parts from them. By using this framework, the user can recognize the action in the text with annotation and select appropriate action parts from the candidate χDLs. Even if the appropriate actions are not included in either of the converted χDLs, the user can easily identify the lack of action information in the χDLs by seeing the annotation.

In section 2, we first describe the existing schema for annotating text. Then, we introduce the user interface of our framework followed by the proposed rule-based conversion method. In section 3, we describe an experiment we conducted to discuss the comparative advantages of our framework by comparing the conversion result of χDLs by CLAIRify and the proposed system. As a result, we confirmed that the proposed system could find action information, which was not feasible by CLAIRify. We also compare the proposed system with SynthReader and discuss the advantages of the proposed system against the existing rule-based system. We found that the proposed method performed better, in terms of finding explicit actions, which were important for the information extraction task, while SynthReader was better at finding implicit actions.

2 Methods

In this section, we first describe an existing schema for visualization. Then, we propose a user interface for human review and an automated conversion method from text to χDL.

2.1 Related work: OSPAR format

We used the OSPAR format[8] for visualization and as an intermediate representation to convert text to χDL. An example of the visualized annotation is shown on the left of Fig. 2. The OSPAR format consists of two tasks: (a) named entity recognition (NER) and (b) relation extraction (RE).

NER is a task to find spans of actions, entities, and parameters. Words that represent actions are annotated as REACTION_STEP. Related entities with actions, such as chemical substances, gas, and instruments are annotated as ENTITY. Labels for representing parameters are TIME, TEMPERATURE, and MODIFIER. MODIFIER is information about parameters other than time and temperature used to perform actions, such as atmosphere, way to add compounds, stirring rate, and others.

RE is a task to find semantic roles between the action and entity/parameter. Semantic roles express the relation between a predicate (verb) and its arguments. In the OSPAR format, semantic roles represented by using PropBank-style semantic roles[32] are used and each usage of a verb has a set of roles called rolesets. There are three labels in the OSPAR format, namely ARG1, ARG2, and ARGM. ARG1 represents the prototypical patient or theme of the verb. ARG2 represents other arguments that depend on rolesets. ARGM represents parameters that do not depend on rolesets.

2.2 User interface

We propose a user interface that consists of three main functions: (a) conversion from text to the OSPAR format by BERT, (b) conversion from the OSPAR format to χDL by rules, and (c) conversion from text to χDL by CLAIRify. Fig. 2 shows the user interface. This editorial process starts by entering an organic synthesis procedure into the text box at the top of the screen.

After the user clicks the "annotate text" button, the text is converted to the OSPAR format and the result is visualized by

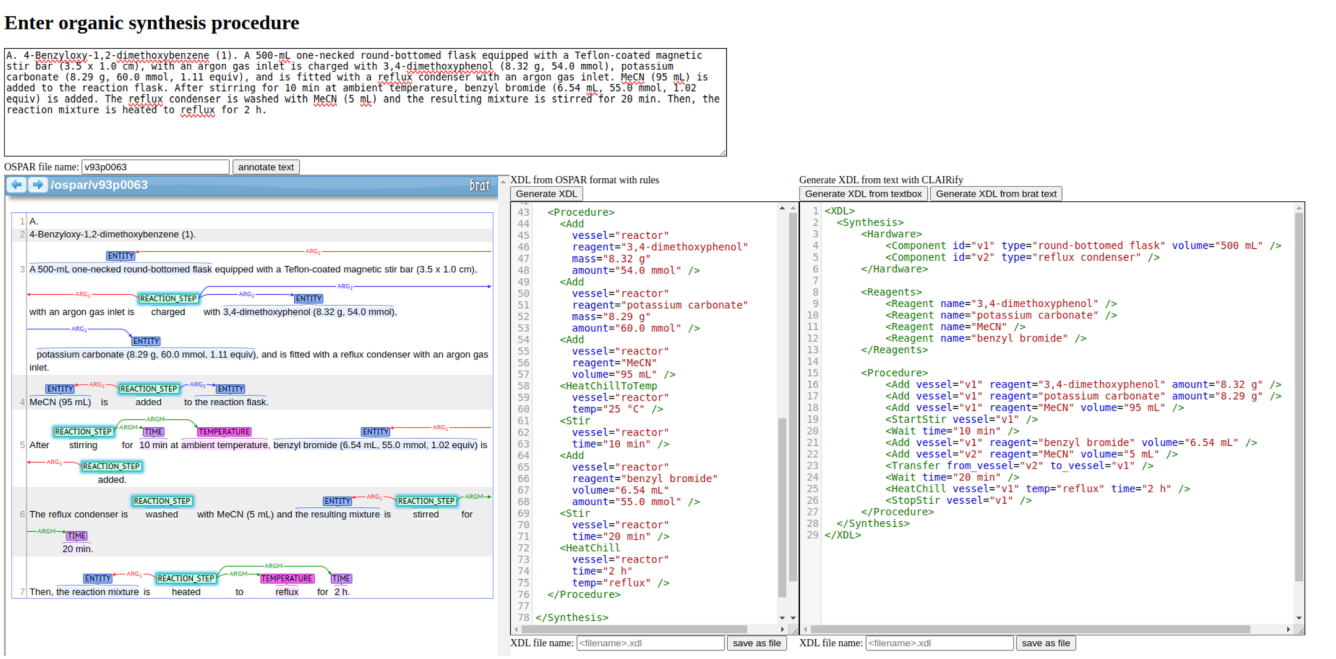


Fig. 2 Screenshot of the proposed user interface. The procedure text is based on Okaya *et al.*,[30] with revisions made through pre-processing in the OSPAR corpus.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

brat,[33] a web-based annotation tool. Then, the user can see the annotated text in the OSPAR format. If the user is not satisfied with the automatic annotation results, he/she can modify the annotation results by using brat as an annotation tool. This modification has the potential to improve the automated conversion method from the OSPAR format to χDL. By moving a cursor over REACTION_STEP, the user can view a roleset for the action to refer to the modification (Fig. S1†).

Then, the user clicks the "Generate χDL" button above the middle text editor to generate χDL from the OSPAR annotation which is displayed in the brat. The generated χDL can be edited and saved by clicking the "save as file" button as a filename beside the button. The text editor on the right is used for conversion from the text in the top textbox or text shown in brat to χDL by CLAIRify. The buttons around this editor have the same functions as the middle text editor. The user can compare both conversion results and select a better χDL as a base for the reviewing process. When the user finds some mistakes in the base χDL, the user can refer to the other χDL to check the existence of correct action for revising the base χDL.

See ESI† for more details of this user interface.

2.3 Conversion from text to χDL

2.3.1 Conversion from text to the OSPAR format. We used a system that was similar to a system constructed in our previous work by using the OSPAR corpus.[8] We trained ChemBERT models,[22] which are domain-specific BERTs for chemistry and for NER, and another ChemBERT model for RE by using the training set and development set of the corpus. The ChemBERT models were implemented using HuggingFace Transformers.[34]

2.3.2 Conversion from the OSPAR format to χDL. We converted the OSPAR format to χDL by mapping the arguments of roleset to χDL actions. We used χDL version 2.0.0.[35] Fig. 3 shows an example of the conversion. First, we defined candidate χDL actions for each roleset to map the rolesets to the χDL actions. Then, to align the roleset arguments with χDL arguments, we specified the type and whether each roleset argument was "required" by referring to the target χDL arguments. For liquid and solid handling actions of χDL such as Add and

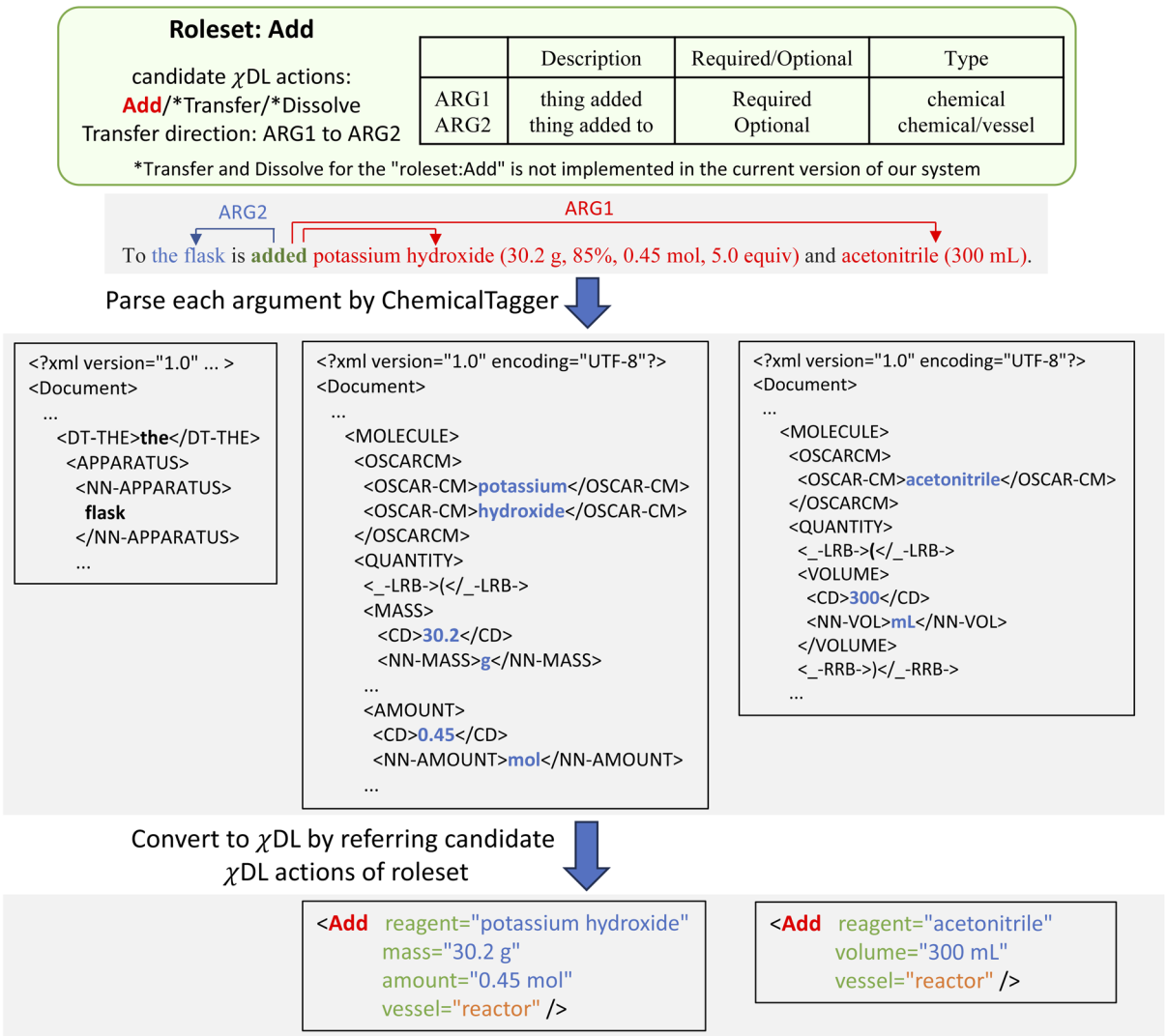


Fig. 3 An example of conversion from the OSPAR format to χDL.

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

Transfer, the corresponding roleset has "transfer direction" information to determine the order of the χDL actions. We used ChemicalTagger[3] to obtain detailed information about the arguments, such as compounds, masses, and other parameters. Then, χDL actions were generated by referring to the roleset and the outputs of ChemicalTagger. When ARGM arguments such as TEMPERATURE or TIME were present, these parameters were either used to generate new actions (*e.g.*, HeatChillToTemp) or to define specific χDL arguments (*e.g.*, time = "1 h"). In cases where an argument was a mixture, such as a solution of sodium iodide (15.0 g, 100 mmol) in acetonitrile (100 mL) (see ESI† for details), a single argument could be converted into multiple χDL actions. The χDL actions were generated as instances of the χDL library, which automatically validated the χDL arguments. Finally, these instances were converted into XML format.

We used text normalization tools. To normalize the verbs, we used WordNet lemmatizer.[36] To interpret numbers in word form, we used text2num[37] to convert numbers in word form to numerical form. Additionally, we defined several constants to accurately interpret the parameters expressed in words and convert them into precise values (see ESI†).

Because it was difficult to capture multiple flasks (*e.g.*, compounds A and B were mixed in a flask X and compounds C + D were mixed in a flask Y), we fixed vessel to reactor other than the case that OSPAR argument was a mixture and multiple compounds were written in the argument.

3 Experiment

As we mentioned in section 1, recall is important for human review. To discuss the advantages of showing both the outputs of the proposed method and CLAIRify, we constructed six χDL examples and evaluated the recall of the methods on the examples. In addition, we compared the proposed method with SynthReader[10] as an existing rule-based method.

3.1 Construction of evaluation data

We selected organic synthesis procedures that can be performed by an automated robot (Chemspeed platform), from *Organic Syntheses*[38] for the examples. The χDL examples were constructed by three authors: two organic chemists (associate and assistant professors) and one information scientist (PhD student). To evaluate the effect of the automated annotation quality on the OSPAR format, we required annotated texts that were checked by humans. Because there were only four examples in the test set of the OSPAR corpus, we additionally annotated two examples other than the corpus in the same manner as the OSPAR corpus including text preprocessing.

To evaluate each method by considering only actions that were explicit in the text, we labeled implicit actions. We labeled each action into an explicit or implicit action to distinguish them in the evaluation phase because errors from these types of actions had different meanings in an information extraction task. There are two types of implicit actions: (a) initiating stirring after addition of reagents and solvents (StartStir) and (b) stirring for a certain period of time (Stir). We did not consider creating a mixture in a noun phrase such as a solution of sodium iodide (15.0 g, 100 mmol) in acetonitrile (100 mL) because the actions were embedded in the text, unlike the abovementioned actions. When creating the correct χDL, if stirring continues after a Stir, it is treated as an implicit action. Therefore, instead of setting continue_stirring=True as an argument for Stir, it was represented as Stir and StartStir, with StartStir being treated as an implicit action. While we annotated the stopping stirring (StopStir) or heating (StopHeatChill) when the target vessels were not used in subsequent steps, we excluded these actions from the evaluation as they were not critical to reproduce the procedure.

3.2 Experimental settings

We compared four methods in this experiment: SynthReader, the proposed method from text to χDL (Pipeline), the proposed method from human-annotated OSPAR format to χDL (OSPAR2χDL), and CLAIRify. To compare the proposed method with/without human intervention, we used Pipeline and OSPAR2χDL. We used SynthReader *via* a web interface called ChemIDE.[39] We used CLAIRify downloaded from github.[40] While the original CLAIRify used GPT-3.5, we used GPT-4o (gpt-4o-2024-05-13) as an LLM because later models were considered to be better than older models. Because the original implementation did not work on the current version of OpenAI API, we made a minor revision of the source code.

In addition to evaluating each system individually, the combination of CLAIRify with other systems for a practical situation by human review was also examined. The combined recall was calculated by verifying whether the correct answer was present among the χDL actions produced by independently running the two systems. To evaluate close failures, we defined action recall in addition to exact recall. The definitions were the following:

- Exact recall: the proportion of correct actions with only correct parameters among the actions in gold data.

Gold data

```
<Add vessel="reactor"
     reagent="water"
     volume="1.0 mL" />
```

	Exact recall	Action recall
`<Add vessel="reactor" reagent="water" volume="1.0 mL" />`	○	○
`<Add vessel="reactor" reagent="water" />`	× (missing parameter)	○
`<Add vessel="reactor" reagent="water" volume="2.0 mL" />`	× (wrong parameter)	○

Fig. 4 An example for evaluating a correct action in exact recall and action recall. While missing and/or wrong parameters are not allowed in exact recall, they are allowed in action recall.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

• Action recall: the proportion of correct actions with correct parameters and correct actions with missing/wrong parameters among the actions in gold data.

An example for evaluating a correct action in exact recall and action recall are shown in Fig. 4.

Other evaluation criteria were the following:

• Liquid/solid handling actions: it was considered as correct action if either mass or volume was specified in an action even if both mass/volume and amount were mentioned in the text. It was acceptable if dropwise=True was not specified for the Add action. It was also acceptable to create the initial mixture in the reactor.

• Stirring actions: when mass or volume were mentioned multiple times in the text, it was considered correct if either mass or volume was specified in an action.

• Temperature control actions: for HeatChill, cases like "between −15 °C and −5 °C" were considered correct if the temperature was set within that range.

3.3 Results and discussion

3.3.1 Evaluation for explicit actions. Table 1 shows the result of explicit actions. To see individual systems, CLAIRify showed the best performance in both exact and action recall. While CLAIRify demonstrated a high action recall (60/65), its exact recall was relatively modest (38/65). This was because that CLAIRify sometimes failed to extract the units of the parameters or even the parameters themselves. We found that when CLAIRify failed to extract the parameters, CLAIRify consistently did not extract the parameters in the example procedure (Fig. 5). Such characteristics were observed in two of the six procedures.

The proposed Pipeline and OSPAR2χDL showed higher recall than SynthReader in both exact and action recalls. While the evaluation data were small, we confirmed that the OSPAR format could represent enough information and the proposed rule-based conversion was better than SynthReader. A major reason why Pipeline was better than SynthReader was because Pipeline employed BERT-based NER and RE in contrast to rule-based SynthReader, which could not extract entities and relations absent from its templates. As a result, we observed differences in the recall of liquid handling actions such as Add and Transfer, which require recognizing compound names (Table S4†).

We confirmed that the modification of the OSPAR annotation could improve the conversion quality because actions, which were not extracted by Pipeline, were sometimes found by OSPAR2χDL, in terms of both exact and action recall. The main reason for the difference is that ChemBERT sometimes failed to extract compounds. In addition, errors in identifying entity boundaries led to missing parameters. As a result, the recall of liquid handling actions of Pipeline was lower than that of OSPAR2χDL (Table S4†). For examples of these errors, see ESI (Fig. S3†). To improve the information extraction system from the text to the OSPAR format, for example, increasing training data and improving a deep learning-based model were required. If the user annotates procedures for χDL conversion, the annotated procedure can be used as training data of the models for text to the OSPAR format.

There were two common errors that were difficult for rule-based methods when converting the OSPAR format to χDL. The first was an incorrect target vessel for actions because the proposed rules could not consider multiple vessels, as described in section 2.3.2. The second was the missing quantity or mass, as ChemicalTagger failed to determine which parameters belonged to which molecule. For example, *o*-tolylboronic acid, 10.0 g (73.6 mmol) was not correctly parsed due to the comma after *o*-tolylboronic acid. While the proposed method was sensitive to the notations of the OSPAR arguments in the text, CLAIRify was robust to these notations thanks to the flexibility of a GLLM.

We also confirmed that combining results of rule-based methods and CLAIRify was effective in increasing exact recalls. To compare the exact recalls of the CLAIRify combined with other systems, Pipeline and OSPAR2χDL showed better results than SynthReader (SynthReader found 7 new actions, Pipeline found 10 actions and OSPAR2χDL found 11 new actions). On the other hand, action recalls of the combined results did not increased by three methods. This result indicates the improved capability of CLAIRify for finding actions. Fig. 5 shows an example of how other systems can improve CLAIRify's recall. In this example, the amount or volume of reagents was not specified by CLAIRify. In other cases, CLAIRify failed to extract the time of addition or the stirring rate, even though these parameters were clearly stated in the text.

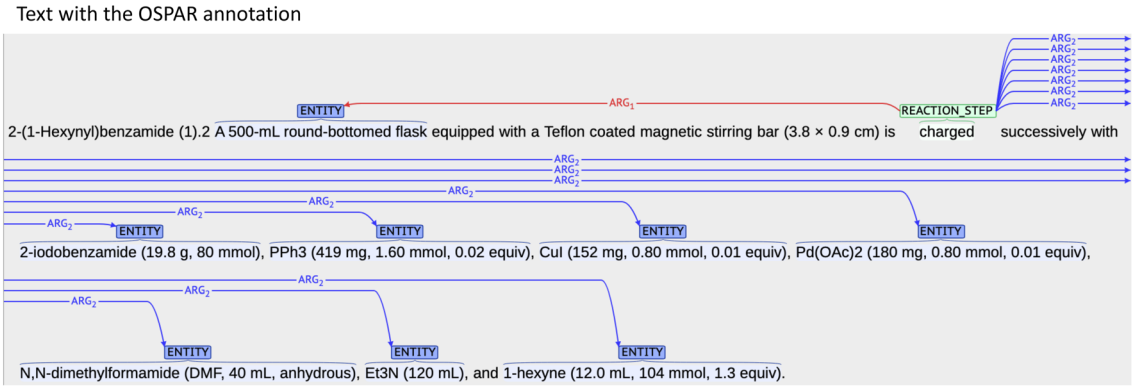
3.3.2 Evaluation for implicit actions. Table 2 shows the result of implicit actions. As we expected, the proposed Pipeline and OSPAR2χDL could not find implicit actions because the OSPAR format did not consider actions that were not written in text and we did not make rules to complement such actions when converting the OSPAR format to χDL. SynthReader could find implicit actions because rules to complement such actions were included in SynthReader. We confirmed that CLAIRify could find implicit actions by the capability of a GLLM.

To enable finding implicit actions by the proposed methods, we need to construct rules to capture these actions in future work. For example, inserting StartStir following multiple Add

Table 1 Result of explicit actions. The numbers indicate (#found action)/(#all actions). SR is SynthReader, Pipe is Pipeline, O2X is OSPAR2χDL and CLAIR is CLAIRify

	SR	Pipe	O2X	CLAIR	SR + CLAIR	Pipe + CLAIR	O2X + CLAIR
Exact recall	22/65	28/65	31/65	38/65	45/65	48/65	49/65
Action recall	34/65	41/65	50/65	60/65	60/65	60/65	60/65

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.



χDL by OSPAR2χDL

```
<Procedure>
 <Add  vessel="reactor"
       reagent="2-iodobenzamide"
       mass="19.8 g"
       amount="80 mmol" />
 <Add  vessel="reactor"
       reagent="PPh3"
       mass="0.419 g"
       amount="1.60 mmol" />
 <Add  vessel="reactor"
       reagent="CuI"
       mass="0.152 g"
       amount="0.80 mmol" />
 <Add  vessel="reactor"
       reagent="Pd(OAc)2"
       mass="0.18 g"
       amount="0.80 mmol" />
 <Add  vessel="reactor"
       reagent="N,N-dimethylformamide DMF"
       volume="40 mL" />
 <Add  vessel="reactor"
       reagent="Et3N"
       volume="120 mL" />
 <Add  vessel="reactor"
       reagent="1-hexyne"
       volume="12 mL"
       amount="104 mmol" />
 ...
</Procedure>
```

χDL by CLAIRify

```
<Procedure>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="2-iodobenzamide"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="PPh3"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="CuI"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="Pd(OAc)2"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="DMF"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="Et3N"/>
 <Add  vessel="round_bottom_flask_500mL"
       reagent="1-hexyne"/>
 ...
</Procedure>
```

Fig. 5 An example for increased recall by systems other than CLAIRify when CLAIRify did not extract parameters. The procedure text is based on Okaya *et al.*,[30] with revisions made through pre-processing in the OSPAR corpus.

Table 2 Result of implicit actions. The numbers indicate (#found action)/(#all actions). SR is SynthReader, Pipe is Pipeline, O2X is OSPAR2χDL and CLAIR is CLAIRify

	SR	Pipe	O2X	CLAIR	SR + CLAIR	Pipe + CLAIR	O2X + CLAIR
Exact recall	2/12	0/12	0/12	6/12	7/12	6/12	6/12
Action recall	2/12	0/12	0/12	6/12	7/12	6/12	6/12

actions would be considered. Another example is inserting StopStir when the reactor or flask was not used after the previous action.

3.3.3 What should be considered in the human review process? As we discussed, we found that CLAIRify was generally promising for generating a "base χDL" because it demonstrated higher recalls than other systems. However, the output of the proposed system may be preferred as a base χDL when CLAIRify consistently fails to extract the parameters.

We found that human reviewers need to be careful with vessel parameters when multiple vessels are used in a procedure. When combining multiple candidate χDLs, reviewers also need to pay attention to the parameters. For example, the vessel names should be standardized to match the notation used in one of the χDLs, and the associated component should be declared in the hardware section. Similarly, the reagents section should be updated when a reagent that used in χDL actions is not declared.

To supplement implicit actions, expertise in chemistry is required because it is necessary to determine when stirring is required and how long the stirring should continue. For example, it is difficult for non-experts to supplement implicit

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM. This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

stirring actions after mixing compounds because the stirring time depends on the specific compounds involved.

4 Conclusions

To curate organic synthesis procedures as structured data, we focused on the limitations of the automated information extraction from the literature and proposed a framework for reviewing the results of the extraction. The proposed user interface can visually support human reviewers by highlighting original texts with the OSPAR format. In addition, to improve the quality of the automated conversion, a method to show the generated χDL by our rule-based system and the GLLM-based CLAIRify was developed. In the experiment, we confirmed the comparative advantage of our approach by showing the outputs of both systems to the user. We also confirmed that our system obtained higher recall than SynthReader. In the future, we plan to improve our system by improving the automated annotation quality of the OSPAR format and maintaining rules to obtain both explicit and implicit actions.

Data availability

The code for our framework can be found at **https://github.com/mlmachi/OSPAR_XDL/**. The fine-tuned ChemBERT models, χDLs for the experiments, and detailed evaluation for each χDL action are available at **https://doi.org/10.6084/m9.figshare.27233541**.

Author contributions

KM: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, software, writing – original draft. SA: conceptualization, data curation. YN: conceptualization, data curation, funding acquisition, resources. MY: conceptualization, funding acquisition, resources, supervision, writing – review & editing.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was partially supported by JST-ERATO (JPMJER1903) and the Institute for Chemical Reaction Design and Discovery (ICReDD), which was established by the World Premier International Research Initiative (WPI), MEXT, Japan. Support was also provided by JSPS KAKENHI grant number JP23H03810 and 23K18500, and JST SPRING, grant number JPMJSP2119.

References

1 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, *et al.*, *Chem. Rev.*, 2024, **124**, 9633–9732.
2 M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, *Sci. Data*, 2016, **3**, 1–9.
3 L. Hawizy, D. M. Jessop, N. Adams and P. Murray-Rust, *J. Cheminf.*, 2011, **3**, 1–13.
4 M. C. Swain and J. M. Cole, *J. Chem. Inf. Model.*, 2016, **56**, 1894–1904.
5 D. Q. Nguyen, Z. Zhai, H. Yoshikawa, B. Fang, C. Druckenbrodt, C. Thorne, R. Hoessel, S. A. Akhondi, T. Cohn, T. Baldwin and K. Verspoor, *Advances in Information Retrieval*, Cham, 2020, pp. 572–579.
6 A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller and T. Laino, *Nat. Commun.*, 2020, **11**, 3601.
7 S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen and C. W. Coley, *J. Am. Chem. Soc.*, 2021, **143**, 18820–18826.
8 K. Machi, S. Akiyama, Y. Nagata and M. Yoshioka, *J. Chem. Inf. Model.*, 2023, **63**, 6619–6628.
9 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *et al.*, *Science*, 2019, **365**, eaax1566.
10 S. H. M. Mehr, M. Craven, A. I. Leonov, G. Keenan and L. Cronin, *Science*, 2020, **370**, 101–108.
11 N. Yoshikawa, M. Skreta, K. Darvish, S. Arellano-Rubach, Z. Ji, L. Bjørn Kristensen, A. Z. Li, Y. Zhao, H. Xu, A. Kuramshin, *et al.*, *Aut. Robots*, 2023, **47**, 1057–1086.
12 K. Laws, M. Tze-Kiat Ng, A. Sharma, Y. Jiang, A. J. Hammer and L. Cronin, *ChemElectroChem*, 2024, **11**, e202300532.
13 E. M. Zamora and P. E. Blower, *J. Chem. Inf. Comput. Sci.*, 1984, **24**, 176–181.
14 E. M. Zamora and P. E. Blower, *J. Chem. Inf. Comput. Sci.*, 1984, **24**, 181–188.
15 N. Kemp and M. Lynch, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 544–551.
16 L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin and J. Wang, *Bioinformatics*, 2018, **34**, 1381–1388.
17 P. Corbett and J. Boyle, *J. Cheminf.*, 2018, **10**, 59.
18 J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, 2019, pp. 4171–4186.
19 I. Beltagy, K. Lo and A. Cohan, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3615–3620.
20 J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So and J. Kang, *Bioinformatics*, 2020, **36**, 1234–1240.
21 Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao and H. Poon, *ACM Transactions on Computing for Healthcare*, 2021, **3**, 1–23.
22 J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen and R. Barzilay, *J. Chem. Inf. Model.*, 2021, **62**, 2035–2045.
23 T. Gupta, M. Zaki, N. A. Krishnan and Mausam, *npj Comput. Mater.*, 2022, **8**, 102.

© 2025 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 27 November 2024. Downloaded on 5/9/2026 12:37:30 PM.
This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

24 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Adv. Neural Inf. Process. Syst.*, 2020, 1877–1901.
25 H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, *arXiv*, 2023, preprint, arXiv:2302.13971, DOI: **10.48550/arXiv.2302.13971**.
26 A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, *J. Mach. Learn. Res.*, 2023, **24**, 1–113.
27 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *Nat. Mach. Intell.*, 2024, 1–11.
28 Q. Ai, F. Meng, J. Shi, B. Pelkie and C. W. Coley, *Digital Discovery*, 2024, **3**, 1822–1831.
29 J. Dagdelen, A. Dunn, S. Lee, N. Walker, A. S. Rosen, G. Ceder, K. A. Persson and A. Jain, *Nat. Commun.*, 2024, **15**, 1418.
30 S. Okaya, K. Okuyama, K. Okano and H. Tokuyama, *Org. Synth.*, 2003, **93**, 63–74.
31 T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, *Adv. Neural Inf. Process. Syst.*, 2020, 1877–1901.
32 M. Palmer, D. Gildea and P. Kingsbury, *Comput. Ling.*, 2005, **31**, 71–106.
33 P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou and J. Tsujii, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France, 2012, pp. 102–107.
34 T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020, pp. 38–45.
35 XDL, **https://gitlab.com/croningroup/chemputer/xdl**, accessed May 9, 2024.
36 G. A. Miller, *Commun. ACM*, 1995, **38**, 39–41.
37 text2num, **https://github.com/allo-media/text2num**, accessed September 23, 2024.
38 *Organic Syntheses*, **https://www.orgsyn.org**, accessed October 14, 2021.
39 ChemIDE, **https://croningroup.gitlab.io/chemputer/xdlapp**, accessed September 23, 2024.
40 CLAIRify, **https://github.com/ac-rad/xdl-generation**, accessed March 18, 2024.

© 2025 The Author(s). Published by the Royal Society of Chemistry