AutoTemplate: Enhancing Chemical Reaction Datasets for Machine Learning Applications in Organic Chemistry

Lung-Yi Chen[1] and Yi-Pei Li[1,2*]

[1*]Department of Chemical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei, 10617, Taiwan.

[2]Taiwan International Graduate Program on Sustainable Chemical Science and Technology (TIGP-SCST), No. 128, Sec. 2, Academia Road, Taipei, 11529, Taiwan.

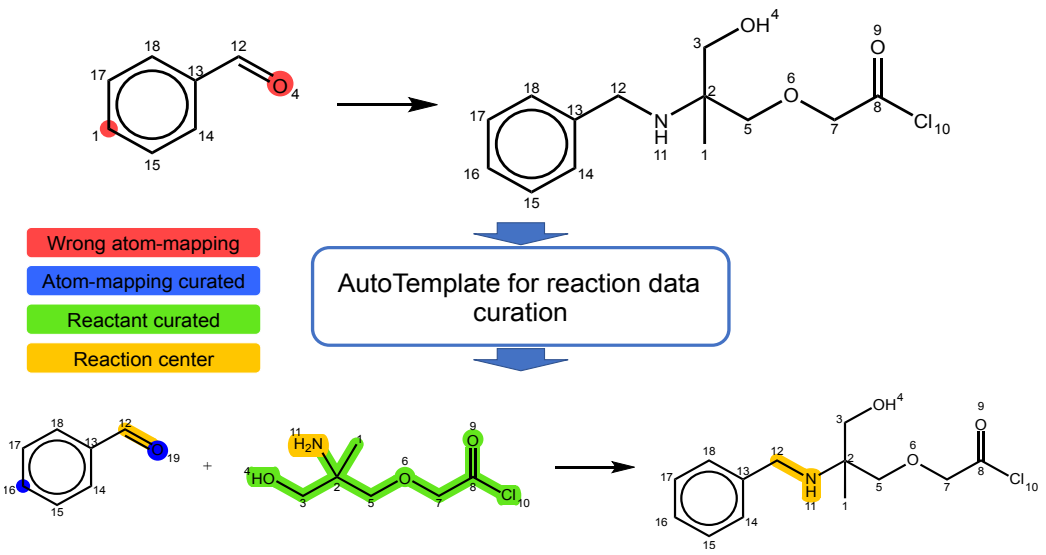
*Corresponding author: Yi-Pei Li. E-mail: yipeili@ntu.edu.tw;

Abstract

This paper presents AutoTemplate, an innovative data preprocessing protocol, addressing the crucial need for high-quality chemical reaction datasets in the realm of machine learning applications in organic chemistry. Recent advances in artificial intelligence have expanded the application of machine learning in chemistry, particularly in yield prediction, retrosynthesis, and reaction condition prediction. However, the effectiveness of these models hinges on the integrity of chemical reaction datasets, which are often plagued by inconsistencies like missing reactants, incorrect atom mappings, and outright erroneous reactions. AutoTemplate introduces a two-stage approach to refine these datasets. The first stage involves extracting meaningful reaction transformation rules and formulating generic reaction templates using a simplified SMARTS representation. This simplification broadens the applicability of templates across various chemical reactions. The second stage is template-guided reaction verification, where these templates are systematically applied to validate and correct the reaction data. This process effectively amends missing reactant information, rectifies atom-mapping errors, and eliminates incorrect data entries. A standout feature of AutoTemplate is its capability to concurrently identify and correct false chemical reactions. It operates on the premise that most reactions in datasets are accurate, using these as templates to guide the correction of flawed entries. The protocol demonstrates its efficacy across a range of chemical reactions, significantly enhancing dataset quality. This advancement provides a more robust foundation for developing reliable machine learning models in chemistry, thereby improving the accuracy of forward and retrosynthetic predictions. AutoTemplate marks a significant progression in the preprocessing of chemical reaction datasets, bridging a vital gap and facilitating more precise and efficient machine learning applications in organic synthesis. *Scientific contribution*: The proposed automated preprocessing tool for chemical reaction data aims to identify errors within chemical databases. Specifically, if the errors involve atom mapping or the absence of reactant types, corrections can be systematically applied using reaction templates, ultimately elevating the overall quality of the database.

Keywords: Reaction template, Data preprocessing, Atom-to-atom mapping, Reaction data curation

Graphical Abstract



1 Introduction

Recent advancements in artificial intelligence have greatly expanded its applications in the field of chemistry. Machine learning techniques have been integrated into various aspects of organic synthesis, including yield prediction [1–4], forward prediction [5–9], retrosynthesis [10–15] and reaction condition prediction [16–19]. These predictive models rely on extensive and reliable chemical reaction datasets, enabling the development of robust machine learning solutions for real-world scenarios [20–24].

Chemical reaction databases commonly utilized in the literature can be broadly categorized as open-source datasets such as the United States Patent and Trademark Office (USPTO) [25] and open reaction database (ORD) [26], or proprietary datasets like Pistachio [27], Reaxys [28], SciFinder [29], and Spresi [30]. These datasets are compiled through text-mining or manual recording, both of which can introduce errors in the chemical reaction data. Fig. 1 illustrates common data deficiencies observed in chemical databases, including missing reactants, inexplicable extra atoms in products, and even entirely erroneous reactions. Detecting and rectifying these data inconsistencies often require human intervention to ensure the quality of machine learning models.

To address these issues, Gimadiev et al. [31] employed atom-to-atom mapping toolkits [32–35] and the CGRTools [36] python library for preprocessing chemical transformations. They used a condensed graph of reaction (CGR), representing the superposition of the reactants and products, to remove duplicate reactions and balance reaction equations, particularly in cases where

simple reagents like amine and water were unspecified. In contrast, Vaucher et al. [37] developed a transformer-based model [38] to complete reaction equations by filling in missing parts of molecules in partial reactions using a sequence-to-sequence approach. Although the model exhibited versatility in handling retrosynthesis, forward prediction, and data curation tasks, it achieved an accuracy of approximately 30% for exact matches, which may pose limitations in its application for extensive preprocessing of external chemical reaction datasets. More recently, Toniato et al. [39] employed the concept of catastrophic forgetting [40] to monitor the learning progress of molecular transformer [9] during training. Data points with difficulty in learning were assumed to be associated with errors and were subsequently removed from the dataset. However, the extent of data removal using this approach significantly depended on the model used, its learning capacity, and hyperparameter selection, rendering it less deterministic.

To the best of our knowledge, existing data-preprocessing methods have limited capacity to detect and correct false chemical reactions simultaneously. This gap has motivated us to develop an advanced data-preprocessing protocol called AutoTemplate in this work. AutoTemplate establishes clear criteria for identifying and removing erroneous data while effectively recovering missing reactants. It operates under the assumption that the majority of reactions in datasets are correct and uses these reactions as templates to guide the curation of incorrect data. The proposed method can successfully identify incorrect reactions, correct faulty atom mapping, and complete missing reactants, providing a solid foundation for the development of data-driven machine learning models, thereby enhancing the performance of forward and retrosynthetic predictions.

2 Method

The data cleaning methodology presented in this work is divided into two stages: generic template extraction and template-guided reaction verification. In the generic template extraction stage, we first identify meaningful reaction transformation rules within the dataset of interest. These rules are then expressed as generic reaction templates using a simplified version of the SMARTS representation [43]. This simplification ensures that the templates can be applied to a wide range of reactions with the same transformation. In the template-guided reaction verification stage, we leverage the list of generic reaction templates to systematically validate the reaction data. This involves applying retro templates to the product. If the original reactants are indeed a subset of the results obtained through template application, the template-applied outcomes replace the original data. This process effectively rectifies any missing reactant information and simultaneously corrects potential atom-mapping errors. However, in situations where none of the templates match the reaction, indicating

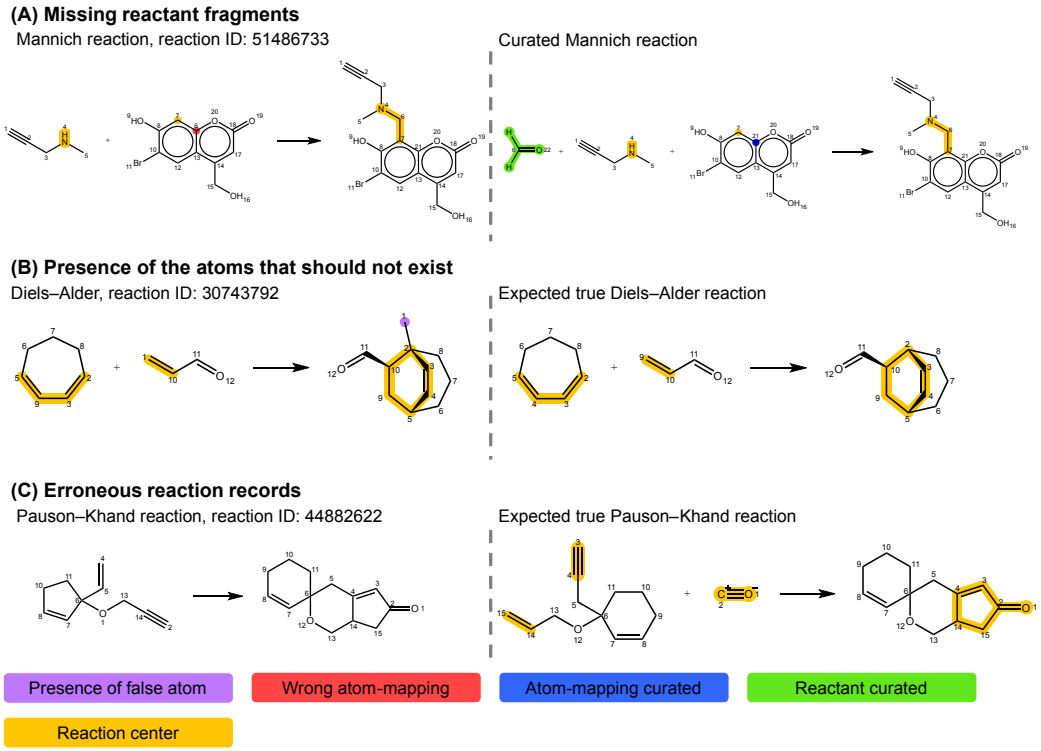


Fig. 1 Illustrations of deficiencies in reaction datasets: (A) The selected Mannich reaction omits formaldehyde in the reactants. (B) The presence of the carbon atom (labeled as purple) violates the law of conservation of matter, and the accurate product based on the work by De Nino et al. [41] is shown on the right. (C) The reactant and product do not match, and the correct chemical reaction is depicted on the right side, extracted from the study by Özdemirhan [42]. These examples are sourced from the Reaxys database [28], but it is important to note that similar errors exist in other databases. Notably, the original Reaxys dataset lacks atom-mapping information, and the atom-mapping labels in the left half of this figure were generated using the RXNMapper software [35].

an unusual chemical transformation and potentially incorrect data entry, we opt to remove that specific reaction from our dataset. The overall procedure is visually depicted in Fig. 2, with detailed step-by-step explanations provided in the following subsections.

2.1 Generic template extraction

2.1.1 Reaction data collection

To evaluate the effectiveness of our data cleaning protocol, we applied it to reaction data derived the Reaxys database [28], a well-established resource in the field of computational chemistry that, like any large database, may contain some errors [31]. To demonstrate the broad applicability of our data preprocessing approach, we retrieved datasets for 20 different reaction types from Reaxys. These datasets were obtained by searching for specific reaction names, and they encompassed a variety of reactions, including Adams decarboxylation, Baylis–Hillman reaction, Buchwald–Hartwig cross coupling, Chan–Lam coupling, Diels–Alder, Fischer indole synthesis, Friedel–Crafts acylation, Friedel–Crafts alkylation, Grignard reaction, Hiyama coupling, Huisgen cycloaddition, Hydrogenation, Kabachnik–Fields reaction, Kumada coupling, Mannich reaction,

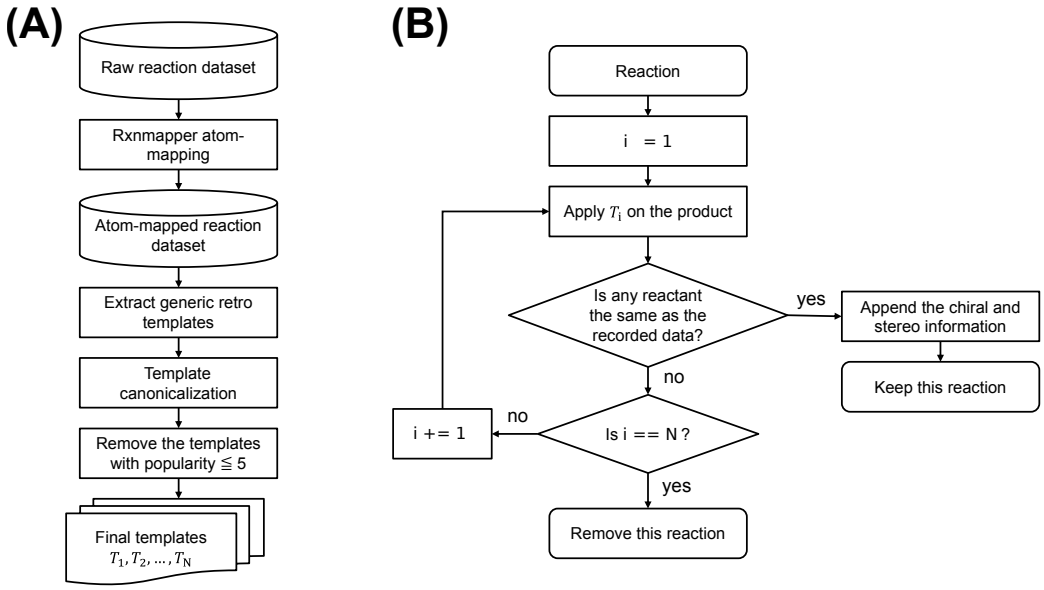


Fig. 2 Overview of the two-stage data cleaning protocol of AutoTemplate for processing chemical reaction data. Panel (A) illustrates the generic template extraction procedure. Panel (B) shows the template-guided reaction verification process, which systematically validates the reaction data using a list of generic reaction templates.

Negishi coupling, Pauson–Khand reaction, reductive amination, Suzuki coupling, and Wittig reaction. The reaction IDs for each reaction used in our study are provided in the GitHub repository for reference [44]. We removed any reactions involving reactants or products that could not be parsed by RDKit [45]. In addition, we eliminated isotope labels from the molecules since they do not impact the chemical transformation. It is worth noting that the labels denoting reaction types in the Reaxys database may not always align accurately with the actual reaction types. Therefore, despite our efforts to collect data based on the 20 specified reaction names, there were instances where the recorded reaction entries did not correspond precisely to these 20 designated reaction types.

2.1.2 Atom-to-atom mapping

The original reaction data obtained from Reaxys lacked information on atom mapping, a crucial element for establishing correspondence between the atoms of reactants and products. This information is essential to identify the reaction center where the connectivity of atoms has changed, a prerequisite for extracting the reaction template. The accuracy of common atom-to-atom mapping toolkits has been assessed in the study by Lin et al. [33]. According to their findings, the open-source tool RXNMapper [35] demonstrated state-of-the-art performance, processing each reaction within one second. Due to these advantages, we selected RXNMapper as our preprocessing toolkit for atom-to-atom mapping. With atom-mapping information available, we can distinguish spectator molecules—those

that do not actively participate in the reaction or contribute any non-hydrogen atoms to the product. These spectator molecules were removed because our data preprocessing framework focuses on curating the chemical transformation itself, rather than the spectator molecules.

2.1.3 Generic template definition and extraction

Upon obtaining the atom-mapped reactions, the next step is to retrieve all the reaction templates from the dataset using the RDChiral [46] template extractor. It is important to note that RDChiral primarily focuses on generating retrosynthetic templates, which are designed for developing computer-aided retrosynthesis models. Because chemical reaction datasets often focus on the major product while not necessarily comprehensively documenting the reactants needed to produce that product, our study utilizes retrosynthetic templates to verify and curate the reaction data.

The default templates generated by RDChiral provide highly detailed information around the reaction center. This results in an excessive number of templates for the same type of chemical transformation, particularly when there are minor variations in neighboring functional groups. It also extends the time required for the subsequent template application process. The specificity of these templates can make it challenging to apply a template from one reaction entry to curate another entry, unless both entries have identical neighboring functional groups near the reaction center. To overcome these challenges, we made modifications to the RDChiral functions. Our aim was to create generic reaction templates that include only essential information concerning atom types and bond types within the reaction centers, while excluding extraneous details. Table 1 provides a comparison between the default and modified template extraction functions.

Consider the Grignard reaction in Fig. 3A as an example, the corresponding reaction template generated by default RDChiral is `[OH;D1;+0:4]-[CH;D3;+0:5](-[c:6])-[c;H0;D3;+0:1](:[c:2]):[c:3]>>Br-[c;H0;D3;+0:1](:[c:2]):[c:3].[O;H0;D1;+0:4]=[CH;D2;+0:5]-[c:6]`. On the other hand, its generic template reduces to `[#6:1]-[#6:2]-[#8:3]>>Br-[#6:1].[#6:2]=[#8:3]`. In the generic template, details related to atomic aromaticity, degree of freedom, number of hydrogen atoms, charge, and extra atoms are all discarded. The meanings of the notations used in the template can be found in the reaction SMARTS documentation [47]. This simplification effectively documents the chemical transformation for most cases. Nevertheless, there are special cases that require unique treatment. The first exception involves specifying the number of connected hydrogens in the generic template to accurately represent species involved in radical reactions, as shown in Fig. 3B. The second exception is the inclusion of the number of charges in the template when the reaction involves charge transfer, as illustrated in Fig. 3C. The third exceptional case

Table 1 The features specified in default RDChiral and generic reaction templates.

Level	Features	RDChiral	Generic
Atom	Reactant radius[1]	1	0
	Product radius[1]	0	0
	Aliphatic or aromatic	Yes	No
	Degree of freedom[2]	Yes	No
	Chirality	Yes	No
	No. of hydrogen atoms	Yes	No, except for radical reactions
	Charge	Yes	No, except for charge transfer reactions
Bond	Bond type	Yes	Yes
	Cis-trans isomerism	Yes	No
Functional groups	Leaving groups	Yes	Yes
	Predefined groups	Yes	No

[1]Radius denotes the extending distance of the neighbor atoms around the reaction center.
[2]Degree of freedom here represents the number of connecting non-hydrogen atoms.

arises when separate reaction centers occur in the product (Fig. 3D). In such cases, the connecting atoms between the reaction centers should be incorporated into the generic template. These connecting atoms can be identified using Dijkstra's algorithm [48], which finds the shortest path between given nodes. This approach ensures that no redundant atoms are included in the template and is effectively applicable to extracting templates for ring-opening reactions.

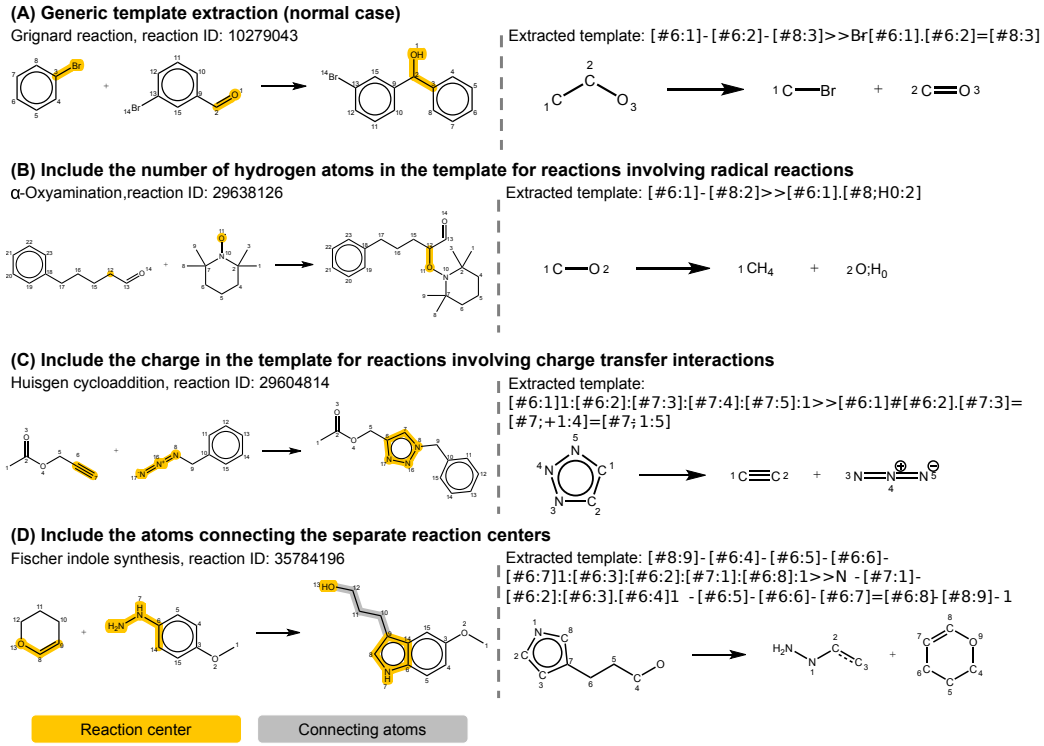


Fig. 3 Illustration of generic template extraction with the normal and special cases.

2.1.4 Template canonicalization

To address the issue of having multiple generic reaction templates representing the same chemical transformation but with different text representations, we employed a graph isomorphism check to confirm whether the reactants and products in pairwise templates were identical. If both reactant and product SMARTS patterns were graph isomorphic, we combined the two templates. Additionally, we calculated the number of bond changes in the templates and keep the one with fewer changes. Fig. 4 illustrates this scenario with two Diels–Alder reaction templates that share identical subgraphs of reactants and products but differ in reaction transformations due to mapping errors from the atom-mapping tool. Such errors can lead to incorrect atom swaps, resulting in additional and incorrect formation and breaking of chemical bonds. Therefore, we retained the template with fewer bond changes.

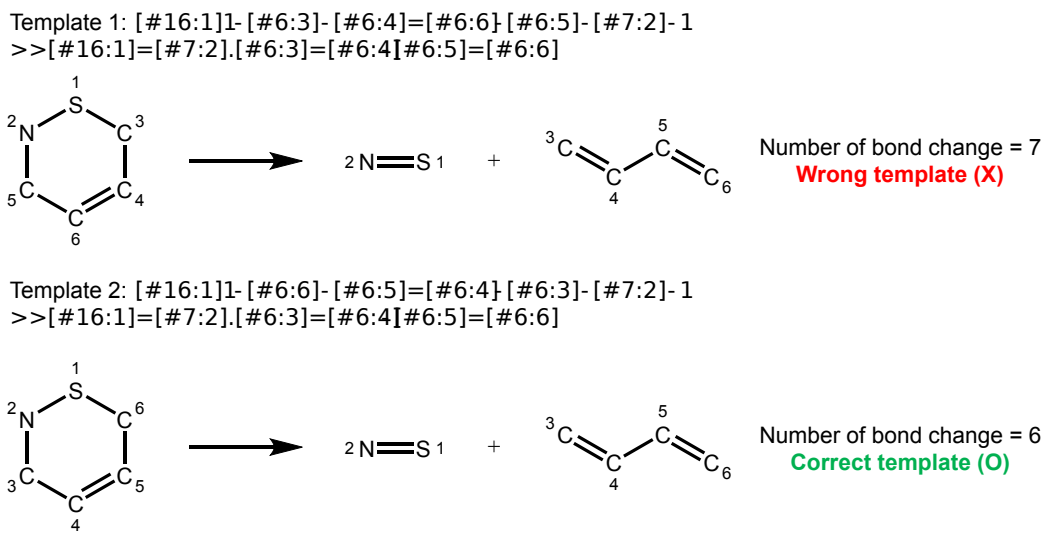


Fig. 4 Examples of two generic templates extracted from Diels–Alder reactions.

2.1.5 Removal of rare templates

Generic templates are designed to be broadly applicable to reaction instances with similar chemical transformations. If a generic template matches only a few reaction entries, it suggests an unusual chemical transformation, possibly indicating that the template may have been derived from a reaction entry with errors. To address this, we monitored the occurrence frequency of each generic template during the template extraction process. Templates with a popularity of 5 or less were removed. This process resulted in the final set of generic templates $\{T_1, T_2, \cdots T_N\}$ for subsequent template-guided verification.

2.2 Template-guided verification

2.2.1 Template application procedure

This procedure primarily involves the iterative application of generic reaction templates to the products of each reaction entry. When the reactants in the original data entry form a subset of the reactants resulting from the applied template, we replace the original data's reactants with those from the applied template. This rectifies any missing reactant information and simultaneously corrects potential atom-mapping errors. In cases where none of the templates match the reaction, indicating an unusual chemical transformation and potentially incorrect data entry, we choose to remove that specific reaction entry from the dataset.

Throughout the template application process, the reactants are automatically supplemented with the appropriate number of hydrogen atoms based on their charge state and the number of bonds connected to them. For instance, neutral sulfur atoms are assigned either two or six bonds, resulting in two possible configurations for a neutral sulfur atom with a connected chemical bond, acquiring either one or five hydrogen atoms. Exceptions to this rule only occur when the template explicitly specifies the number of hydrogen atoms connected to the reaction center.

2.2.2 Append atomic chirality and bond stereochemistry

We note that the reactants generated from template application lack annotations for atomic chirality and bond stereochemistry at the reaction centers. Therefore, an additional step is necessary to reintroduce this information into the reactants, but only if this information was included in the original dataset. This process involves establishing a one-to-one atom correspondence between the original reactants and template-generated reactants. This can be achieved by initially converting both sets of reactants into undirected graphs, followed by utilizing the exact graph matching algorithm [49] to establish a strict one-to-one node correspondence between the two graphs.

3 Results and Discussion

3.1 Analysis of overall results

Table 2 provides information on the number of reactions in the dataset, the number of templates extracted from these reactions, and the residual proportion after data processing. The variation in the number of templates for each type of reaction is due to the unique characteristics of their reaction mechanisms. For example, coupling reactions that involve multiple possible leaving groups often result in a higher template count. Conversely, reductive amination, where the carbonyl group

Table 2 The data preprocessing results for the chemical reaction datasets.

Reaction type	No. of reactions	No. of generic templates	Residual rate
Adams decarboxylation	2,636	54	62.3%
Baylis–Hillman reaction	7,507	84	81.3%
Buchwald–Hartwig cross coupling	18,341	96	90.7%
Chan–Lam coupling	6,885	43	92.1%
Diels–Alder	18,757	258	74.8%
Fischer indole synthesis	6,841	28	85.9%
Friedel–Crafts acylation	10,095	118	82.9%
Friedel–Crafts alkylation	17,248	164	81.3%
Grignard reaction	13,530	154	73.2%
Hiyama coupling	4,089	106	81.7%
Huisgen cycloaddition	54,183	144	94.1%
Hydrogenation	41,217	306	69.4%
Kabachnik–Fields reaction	5,575	14	91.4%
Kumada coupling	16,371	82	89.1%
Mannich reaction	29,698	271	86.0%
Negishi coupling	10,909	146	84.9%
Pauson–Khand reaction	2,703	19	72.4%
Reductive amination	50,406	16	97.1%
Suzuki coupling	184,219	216	98.2%
Wittig reaction	16,337	94	84.8%

is reduced to an amine, has a large number of reaction entries, but only 16 reaction templates are extracted, indicating less variation in its reaction transformation.

Fig. 5 displays curated reaction results, addressing issues such as false atom-mapping, reactant omissions, and the identification and removal of incorrect reaction records. Notably, the Diels–Alder reactions exhibited a high atom-mapping correction rate of 29.3%. This is likely attributed to the complexity of Diels–Alder reactions, which involve numerous bond transformations and instances of intramolecular or fused ring formation, making them challenging for accurate atom-mapping predictions. Conversely, coupling reactions generally showed relatively fewer atom-mapping errors, likely because they involve fewer bond changes. Accurate atom-mapping data can significantly improve reaction prediction quality, particularly for graph-based models. Regarding the issue of missing reactants, Fischer indole synthesis, Kabachnik–Fields reaction, Pauson–Khand reaction, and reductive amination display a noteworthy proportion of data with absent reactants. In the case of the Pauson–Khand reaction, most instances systematically omit carbon monoxide as a reactant. However, there is no clear pattern indicating which reactants may be omitted in the data for the other three types of reactions. Further discussions on specific data errors and curated results are provided in the following subsections for selected examples.

3.2 Visualized results of selected mapping curated examples

Currently, there is no package available that can generate atom-mapping information perfectly for all reactions [33]. In this study, the data-driven neural network RXNMapper [35] was utilized to predict atom mapping. However, it is important to note that even for reactions considered relatively

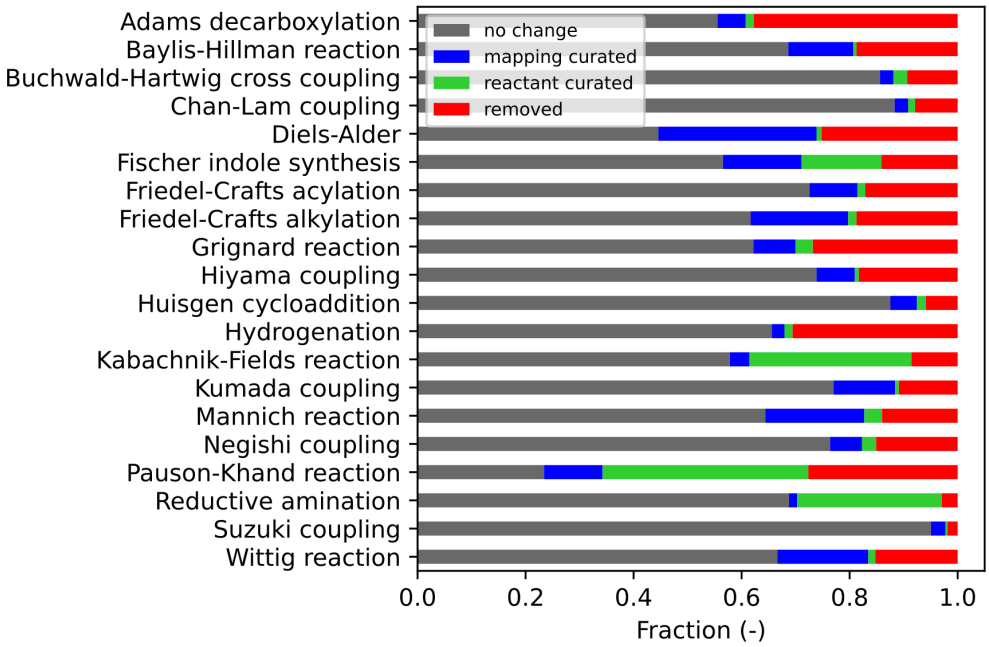


Fig. 5 Distribution of the proportion of repaired reactions after data processing.

straightforward for humans, there can still be instances of incorrect atom mapping, as shown in Fig. 6A. This example of the Baylis-Hillman reaction incorrectly assigns the atom-mapping number (6 and 14) at the position of the carbon-carbon double bond, which would lead to the incorrect reaction template during template extraction. Applying the data processing procedure proposed in this work can recover this reaction with the true atom-mapping labels. Another example is the Buchwald–Hartwig cross-coupling reaction illustrated in Fig. 6B, which has the same issue at the reaction center where the carbon atoms are labeled incorrectly in the intramolecular ring-closing reaction. We note that false atom-mapping issues occur more frequently at the reaction centers, and systematically addressing this problem would benefit downstream template-based and graph-based model applications.

3.3 Visualized results of selected reactant curated examples

The data processing procedure proposed in this work primarily focuses on addressing omitted reactants rather than products, as byproducts and leaving groups are typically not the main focus and are not specified in reaction datasets. The issue of missing reactants can be identified by comparing the atom counts between reactants and products, with reactions having fewer atoms on the reactant side categorized as this type of error. To the best of our knowledge, there is no existing approach tailored for adding missing reactants. However, with the template-guided verification method proposed in this work, erroneous reaction entries can be recovered along with the omitted reactants.

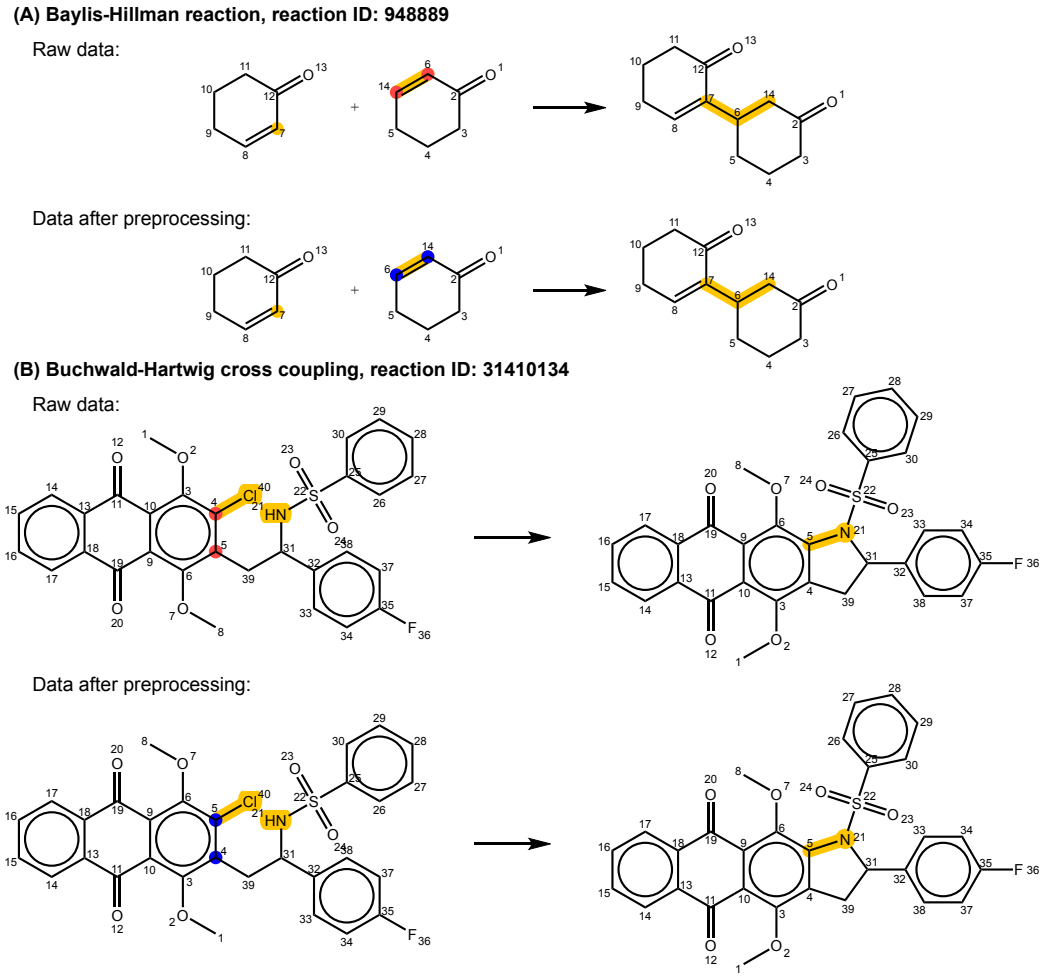


Fig. 6 Two selected examples of (A) Baylis–Hillman reaction and (B) Buchwald–Hartwig cross coupling to demonstrate the curated results of the reaction entries with incorrect atom-mapping. Yellow highlights indicate the reaction centers, red highlights denote atoms with incorrect atom mapping, and blue highlights represent atoms with curated mapping.

Fig. 7A illustrates a typical example from the reductive amination dataset, where the missing reactant with an amine functional group was generated by applying the generic template to the product, thus balancing the reaction equation. In the case of the second instance of the Kabachnik–Fields reaction shown in Fig. 7B, which involves three molecules in the reaction, the two missing fragments were successfully recovered from the template. It is worth noting that the chirality of the phosphorus atom cannot be inferred because the generic template does not specify chiral and cis-trans stereoisomerism at the reaction center. Including such detailed information in templates would lead to an excessive number of templates, reducing the chances of applying a template from one reaction entry to curate another entry.

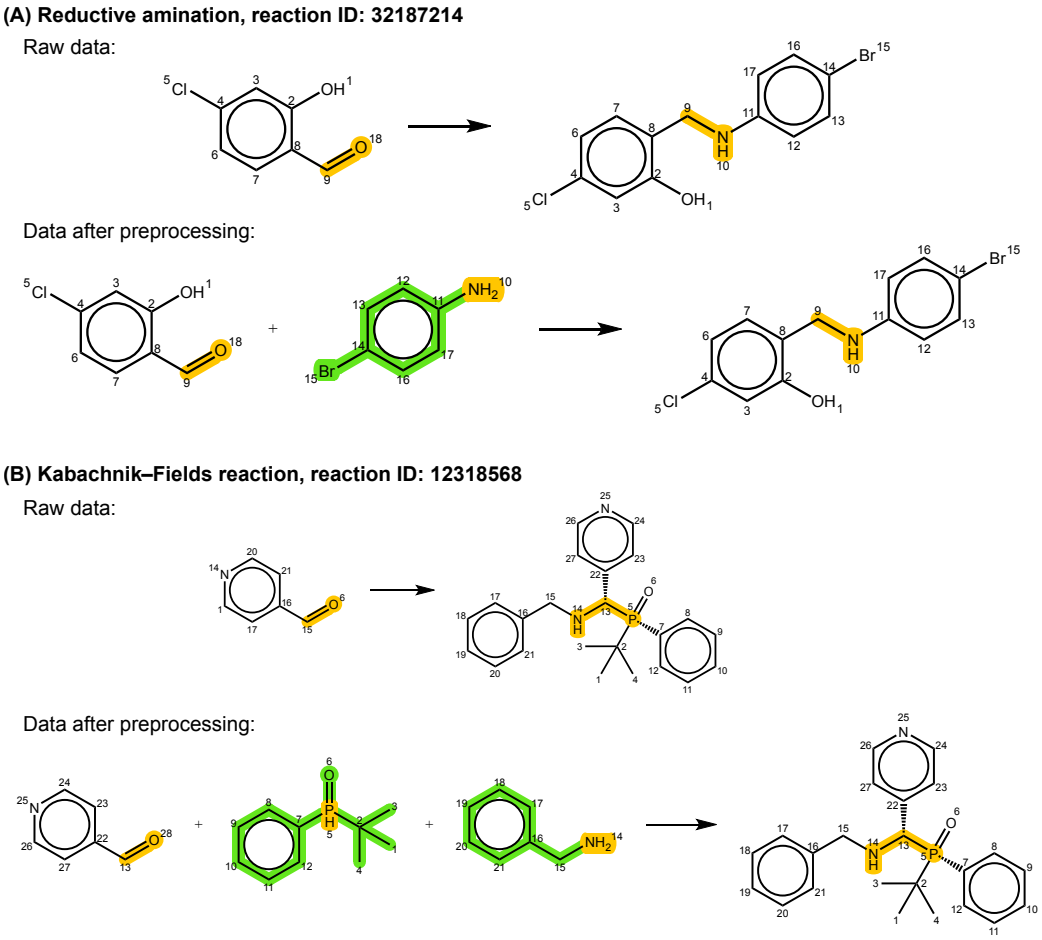


Fig. 7 Two selected examples of (A) reductive amination and (B) Kabachnik–Fields reaction to demonstrate the curated results of the reaction entries with incomplete reactant information. Yellow highlights represent the reaction centers, while green highlights indicate molecular fragments added through the data curation process.

3.4 Visualized results of selected removed reactions

In cases where none of the templates matched the reaction, indicating an unusual chemical transformation or potential data entry errors, the specific reaction entry was removed from the dataset. Several examples of such removals are presented in Fig. 8 and discussed below.

Fig. 8A illustrates a two-step Suzuki coupling reaction. To automatically identify multi-step reactions like this, one would need to repetitively validate them using all the single-step reaction templates, which becomes increasingly time-consuming as the number of steps allowed grows. Because most reaction prediction models focus on single-step reactions, the accommodation of multi-step reactions is less critical in this study. The reactions shown in Fig. 8B and 8C are actually correct reactions, but none of the generic templates in the final list match them. This occurred because the templates extracted from these reactions did not match a sufficient number of reaction entries, leading to their exclusion from the final list of generic templates. As discussed in the method section, templates with low matching frequencies may indicate errors in the template source data. While

this approach effectively removes erroneous reaction entries, it can also inadvertently exclude rare but valid reactions, as demonstrated in Fig. 8B and 8C. The reaction depicted in Fig. 8D belongs to the category of Huisgen cycloaddition. In this reaction, the atom highlighted in purple (number 10) in the product is identified as a carbon atom. However, at the same position in the reactant, an oxygen atom is indicated. Rectifying this type of error is challenging because it is difficult to determine whether the correct structure should be attributed to the reactant or the product. This particular entry originates from a study by McNitt et al. [50], where atom number 10 was labeled as an oxygen atom, suggesting a potential error in the recorded product information in the database.

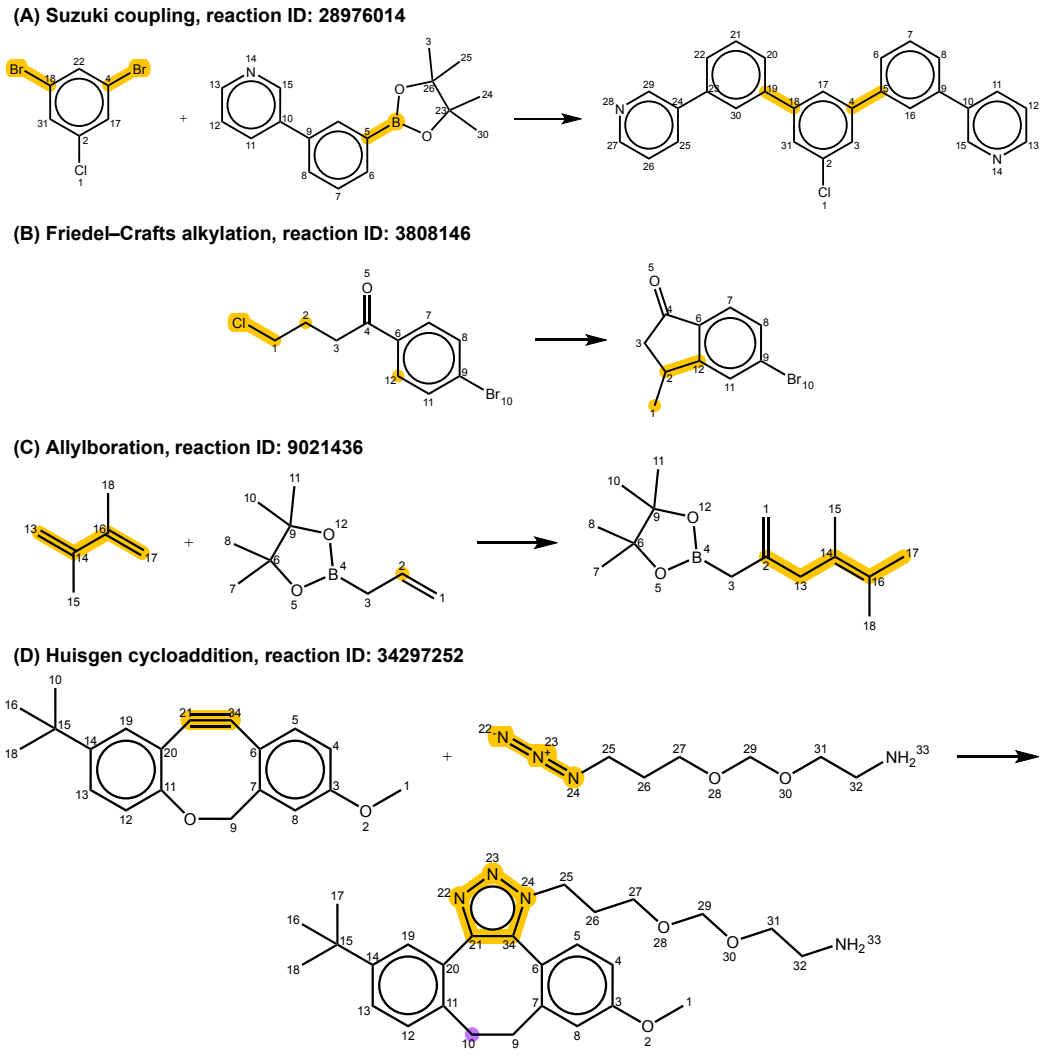


Fig. 8 Four selected examples of (A) Suzuki coupling, (B) Friedel–Crafts alkylation, (C) allylboration, and (D) Huisgen cycloaddition to illustrate reactions that did not match any of the final generic templates and were consequently removed during the data processing procedure.

4 Conclusions

Recent advancements in artificial intelligence have significantly impacted the field of organic chemistry. The reliability of predictive models in chemistry, essential for applications such as yield prediction, retrosynthesis, and reaction condition prediction, is heavily contingent on the quality of chemical reaction datasets. However, these datasets, sourced from both open-source and proprietary databases, often contain inconsistencies like missing reactants, incorrect atom mappings, or erroneous reactions, necessitating rigorous data preprocessing.

This work introduces a novel data preprocessing protocol called AutoTemplate, designed to enhance the quality of chemical reaction datasets. AutoTemplate employs a two-stage approach: generic template extraction and template-guided reaction verification. The process begins with the extraction of meaningful reaction transformation rules from a dataset, which are then expressed as generic reaction templates using a simplified version of the SMARTS representation. This simplification ensures broad applicability across various reactions. In the subsequent stage, these generic templates are systematically applied to validate and correct reaction data. This involves rectifying missing reactant information, correcting atom-mapping errors, and removing incorrect data entries.

Our method stands out by its ability to simultaneously identify and correct false chemical reactions, leveraging the assumption that the majority of reactions in datasets are correct. By using these reactions as templates for data curation, AutoTemplate not only rectifies existing errors but also aids in the recovery of missing reactants. The protocol's effectiveness is demonstrated through its application to diverse chemical reactions, highlighting significant improvements in dataset quality. This refined data provides a more reliable foundation for developing machine learning models in chemistry, enhancing the accuracy of forward and retrosynthetic predictions.

This study represents a significant step forward in preprocessing chemical reaction datasets, addressing a critical gap in the field and paving the way for more accurate and efficient machine learning applications in organic synthesis.

Abbreviations

CGR: Condensed Graph of Reaction; ORD: Open Reaction Database; SMILES: Simplified Molecular Input Line Entry Specification; SMARTS: SMILES Arbitrary Target Specification; USPTO: United States Patent and Trademark Office

Declarations

Availability of data and materials

Full code and reaction IDs for searching the reactions are available at: https://github.com/Lung-Yi/AutoTemplate

Acknowledgements

We are grateful to the National Center for High-performance Computing (NCHC) and the Computer and Information Networking Center at NTU for the support of computing facilities. AI tools were utilized in the process of correcting grammatical mistakes and enhancing the fluency of the manuscript.

Competing interests

The authors declare no competing financial interest.

Funding

Y.P.L. is supported by Taiwan NSTC Young Scholar Fellowship Einstein Program (112-2636-E-002-005) and the Higher Education Sprout Project by the Ministry of Education in Taiwan (113L891305).

Authors' contributions

LYC: Methodology, Formal Analysis, Writing - Original Draft. YPL: Funding Acquisition, Supervision, Writing - Review & Editing.

References

[1] Jiang, S., Zhang, Z., Zhao, H., Li, J., Yang, Y., Lu, B.-L., Xia, N.: When smiles smiles, practicality judgment and yield prediction of chemical reaction via deep chemical language processing. IEEE Access **9**, 85071–85083 (2021)

[2] Probst, D., Schwaller, P., Reymond, J.-L.: Reaction classification and yield prediction using the differential reaction fingerprint drfp. Digital discovery **1**(2), 91–97 (2022)

[3] Saebi, M., Nan, B., Herr, J.E., Wahlers, J., Guo, Z., Zurański, A.M., Kogej, T., Norrby, P.-O., Doyle, A.G., Chawla, N.V.: On the use of real-world datasets for reaction yield prediction. Chemical Science **14**(19), 4997–5005 (2023)

[4] Schwaller, P., Vaucher, A.C., Laino, T., Reymond, J.-L.: Prediction of chemical reaction yields using deep learning. Machine learning: science and technology **2**(1), 015016 (2021)

[5] Coley, C.W., Barzilay, R., Jaakkola, T.S., Green, W.H., Jensen, K.F.: Prediction of organic reaction outcomes using machine learning. ACS central science **3**(5), 434–443 (2017)

[6] Coley, C.W., Jin, W., Rogers, L., Jamison, T.F., Jaakkola, T.S., Green, W.H., Barzilay, R., Jensen, K.F.: A graph-convolutional neural network model for the prediction of chemical reactivity. Chemical science **10**(2), 370–377 (2019)

[7] Do, K., Tran, T., Venkatesh, S.: Graph transformation policy network for chemical reaction prediction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 750–760 (2019)

[8] Fooshee, D., Mood, A., Gutman, E., Tavakoli, M., Urban, G., Liu, F., Huynh, N., Van Vranken, D., Baldi, P.: Deep learning for chemical reaction prediction. Molecular Systems Design & Engineering **3**(3), 442–452 (2018)

[9] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C.A., Bekas, C., Lee, A.A.: Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS central science **5**(9), 1572–1583 (2019)

[10] Coley, C.W., Green, W.H., Jensen, K.F.: Machine learning in computer-aided synthesis planning. Accounts of chemical research **51**(5), 1281–1289 (2018)

[11] Coley, C.W., Rogers, L., Green, W.H., Jensen, K.F.: Computer-assisted retrosynthesis based on molecular similarity. ACS central science **3**(12), 1237–1245 (2017)

[12] Dong, J., Zhao, M., Liu, Y., Su, Y., Zeng, X.: Deep learning in retrosynthesis planning: datasets, models and tools. Briefings in Bioinformatics **23**(1), 391 (2022)

[13] Schreck, J.S., Coley, C.W., Bishop, K.J.: Learning retrosynthetic planning through simulated experience. ACS central science **5**(6), 970–981 (2019)

[14] Tu, Z., Coley, C.W.: Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. Journal of chemical information and modeling **62**(15), 3503–3513 (2022)

[15] Zhong, W., Yang, Z., Chen, C.Y.-C.: Retrosynthesis prediction using an end-to-end graph

generative architecture for molecular graph editing. Nature Communications **14**(1), 3009 (2023)

[16] Chen, L.-Y., Li, Y.-P.: Enhancing chemical synthesis: a two-stage deep neural network for predicting feasible reaction conditions. Journal of Cheminformatics **16**(1), 1–14 (2024)

[17] Gao, H., Struble, T.J., Coley, C.W., Wang, Y., Green, W.H., Jensen, K.F.: Using machine learning to predict suitable conditions for organic reactions. ACS central science **4**(11), 1465–1476 (2018)

[18] Kwon, Y., Kim, S., Choi, Y.-S., Kang, S.: Generative modeling to predict multiple suitable conditions for chemical reactions. Journal of Chemical Information and Modeling **62**(23), 5952–5960 (2022)

[19] Maser, M.R., Cui, A.Y., Ryou, S., DeLano, T.J., Yue, Y., Reisman, S.E.: Multilabel classification models for the prediction of cross-coupling reaction conditions. Journal of Chemical Information and Modeling **61**(1), 156–166 (2021)

[20] Ahneman, D.T., Estrada, J.G., Lin, S., Dreher, S.D., Doyle, A.G.: Predicting reaction performance in c–n cross-coupling using machine learning. Science **360**(6385), 186–190 (2018)

[21] Chen, Y., Zhang, L.: How much can deep learning improve prediction of the responses to drugs in cancer cell lines? Briefings in bioinformatics **23**(1), 378 (2022)

[22] Li, B., Su, S., Zhu, C., Lin, J., Hu, X., Su, L., Yu, Z., Liao, K., Chen, H.: A deep learning framework for accurate reaction prediction and its application on high-throughput experimentation data. Journal of Cheminformatics **15**(1), 1–12 (2023)

[23] Panteleev, J., Gao, H., Jia, L.: Recent applications of machine learning in medicinal chemistry. Bioorganic & medicinal chemistry letters **28**(17), 2807–2815 (2018)

[24] Chen, L.-Y., Li, Y.-P.: Machine Learning Applications in Chemical Kinetics and Thermochemistry, pp. 203–226. Springer, ??? (2023)

[25] Lowe, D.: Chemical reactions from US patents (1976-Sep2016) (2017). https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 Accessed September 05, 2023

[26] Kearnes, S.M., Maser, M.R., Wleklinski, M., Kast, A., Doyle, A.G., Dreher, S.D., Hawkins, J.M., Jensen, K.F., Coley, C.W.: The open reaction database. Journal of the American Chemical

Society **143**(45), 18820–18826 (2021)

[27] Nextmove Software Pistachio (2023). https://www.nextmovesoftware.com/pistachio.html Accessed September 05, 2023

[28] Reaxys (2023). https://www.reaxys.com/ Accessed September 05, 2023

[29] CAS, SciFinder-n (2023). https://scifinder-n.cas.org/ Accessed September 05, 2023

[30] Roth, D.L.: SPRESIweb 2.1, a selective chemical synthesis and reaction database. ACS Publications (2005)

[31] Gimadiev, T.R., Lin, A., Afonina, V.A., Batyrshin, D., Nugmanov, R.I., Akhmetshin, T., Sidorov, P., Duybankova, N., Verhoeven, J., Wegner, J.: Reaction data curation i: chemical structures and transformations standardization. Molecular Informatics **40**(12), 2100119 (2021)

[32] Chen, W.L., Chen, D.Z., Taylor, K.T.: Automatic reaction mapping and reaction center detection. Wiley Interdisciplinary Reviews: Computational Molecular Science **3**(6), 560–593 (2013)

[33] Lin, A., Dyubankova, N., Madzhidov, T.I., Nugmanov, R.I., Verhoeven, J., Gimadiev, T.R., Afonina, V.A., Ibragimova, Z., Rakhimbekova, A., Sidorov, P.: Atom-to-atom mapping: a benchmarking study of popular mapping algorithms and consensus strategies. Molecular Informatics **41**(4), 2100138 (2022)

[34] Nugmanov, R., Dyubankova, N., Gedich, A., Wegner, J.K.: Bidirectional graphormer for reactivity understanding: neural network trained to reaction atom-to-atom mapping task. Journal of Chemical Information and Modeling **62**(14), 3307–3315 (2022)

[35] Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., Laino, T.: Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Science Advances **7**(15), 4166 (2021)

[36] Nugmanov, R.I., Mukhametgaleev, R.N., Akhmetshin, T., Gimadiev, T.R., Afonina, V.A., Madzhidov, T.I., Varnek, A.: Cgrtools: Python library for molecule, reaction, and condensed graph of reaction processing. Journal of chemical information and modeling **59**(6), 2516–2521 (2019)

[37] Vaucher, A.C., Schwaller, P., Laino, T.: Completion of partial reaction equations. Chemrxiv

(2020)

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)

[39] Toniato, A., Schwaller, P., Cardinale, A., Geluykens, J., Laino, T.: Unassisted noise reduction of chemical reaction datasets. Nature Machine Intelligence **3**(6), 485–494 (2021)

[40] Goodfellow, I.J., Mirza, M., Xiao, D., Courville, A., Bengio, Y.: An empirical investigation of catastrophic forgetting in gradient-based neural networks. arXiv preprint arXiv:1312.6211 (2013)

[41] De Nino, A., Bortolini, O., Maiuolo, L., Garofalo, A., Russo, B., Sindona, G.: A sustainable procedure for highly enantioselective organocatalyzed diels–alder cycloadditions in homogeneous ionic liquid/water phase. Tetrahedron letters **52**(13), 1415–1417 (2011)

[42] Özdemirhan, D.: Optically active tertiary alcohols by biocatalysis. Synthetic Communications **47**(7), 629–645 (2017)

[43] Dolfus, U., Briem, H., Rarey, M.: Visualizing generic reaction patterns. Journal of Chemical Information and Modeling **62**(19), 4680–4689 (2022)

[44] Chen, L.-Y.: AutoTemplate. https://github.com/Lung-Yi/AutoTemplate Accessed September 05, 2023

[45] RDKit: Open-Source Cheminformatics Software. http://www.rdkit.org/ Accessed September 05, 2023

[46] Coley, C.W., Green, W.H., Jensen, K.F.: Rdchiral: An rdkit wrapper for handling stereochemistry in retrosynthetic template extraction and application. Journal of chemical information and modeling **59**(6), 2529–2537 (2019)

[47] Daylight SMARTS Documentation. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html Accessed September 05, 2023

[48] Dijkstra, E.W.: A note on two problems in connexion with graphs. In: Edsger Wybe Dijkstra: His Life, Work, and Legacy, pp. 287–290 (2022)

[49] Riesen, K., Jiang, X., Bunke, H.: Exact and inexact graph matching: Methodology and

applications. Managing and mining graph data, 217–247 (2010)

[50] McNitt, C.D., Popik, V.V.: Photochemical generation of oxa-dibenzocyclooctyne (odibo) for metal-free click ligations. Organic & biomolecular chemistry **10**(41), 8200–8202 (2012)