Automatic Generation of Mechanistic Pathways of Organic Reactions with Dual Templates

Anonymous Author(s)
Affiliation
Address
email

Abstract

Understanding organic reaction mechanisms is crucial for interpreting the formation of products at the atomic and electronic level, but still remains as a domain of knowledgeable experts. The lack of a large-scale dataset with chemically reasonable mechanistic sequences also hinders the development of reliable machine learning models to predict organic reactions based on mechanisms as human chemists do. Here, we propose a method that automatically generates reaction mechanisms of a large dataset of organic reactions using autonomously extracted reaction templates and expert-coded mechanistic templates. By applying this method, we labeled 94.8% of 33k USPTO reactions into chemically reasonable arrow-pushing diagrams, validated by expert chemists. Our method is simple, flexible, and can be expanded to cover a wider range of reactions, regardless of type or complexity. We envision it becoming an invaluable tool to propose reaction mechanisms, and to develop future reaction outcome prediction models and discover new reactions.

1 Introduction

The ability to predict reaction outcomes for a given set of substrates, known as forward synthesis prediction, plays a crucial role in successful synthetic planning [1, 2]. This encompasses not only the prediction of major products resulting from organic reactions but also the retrosynthetic analysis, which aims to identify a viable synthetic pathway to synthesize a desired target compound [3]. To ensure the reliability of retrosynthetic analysis, it is essential to couple retrosynthesis prediction models with reliable reaction outcome prediction models. While recent advances in machine learning models have shown promise in predicting reaction outcomes based on reactant sets [4, 5, 6, 7, 8, 9, 10], these models often overlook the finer details of electron movements, reactive intermediates, and other mechanistic information that are crucial for a comprehensive understanding of the reaction [11]. Consequently, there is a need for more sophisticated and accurate chemical models that can explicitly capture the underlying reaction mechanisms, which involve a step-by-step sequence of electron movements and reactive intermediates, to gain valuable insights into the stereochemistry, reaction kinetics, formation of byproducts, and other important reaction details.

Arrow pushing diagrams, commonly utilized by organic chemists, provide a visual representation of electron rearrangements as bonds form and break [12]. While an ideal chemical model is expected to predict the same arrow pushing diagrams as human chemists do, a reliable chemical model that can predict arrow pushing diagrams has not been developed yet due to the lack of mechanistic reaction dataset. In an early attempt to automate this process, Chen and Baldi developed Reaction Explorer [13] to predict major products and mechanistic steps based on reactants and reagents utilizing a set of prioritized transformation rules. While it provided detailed and reasonable mechanistic descriptions, due to the nature of rule prioritizing, introducing a new set of rules to cover more reactions would require revision of significant proportion of existing ones. Hence it is challenging to be scaled up to

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

larger reaction datasets containing diverse reaction types. In a more recent effort, Bradshaw et al. proposed a machine learning-based electro path prediction model called ELECTRO [14] which is fast and scalable. However, since the model still uses only the reactant and product information to extract the electron movements without the chemical guidance of actual mechanisms, most of the electro paths predicted by the models are approximate mechanisms. This gives rises to discrepancies between predicted and actual reaction mechanisms, as illustrated in Figure 1 for example, and highlights the pressing need for a more reliable and dependable mechanistic prediction model. Due to these challenges, a deficiency in a comprehensive database for large-scale reaction mechanisms persists.

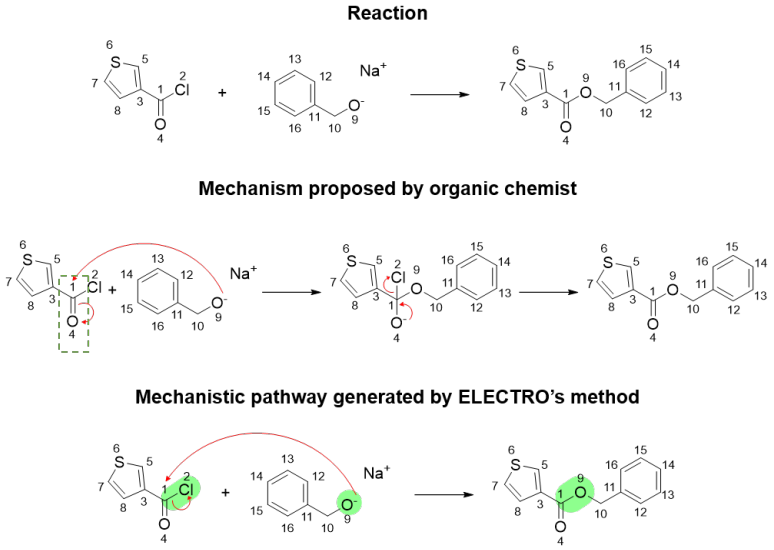

Figure 1: Demonstration of the difference between the more reasonable mechanism intuitively accepted by human chemists and the label obtained by ELECTRO for a nucleophilic acyl substitution reaction between an alkoxide and acyl chloride. Breaking and re-forming of carbonyl bonds (green box) account for the observed chemical reactivity and are reflected in the mechanism. On the other hand, the label obtained by ELECTRO is focused on the changed atoms mainly, highlighted in green: electrophile carbon, leaving group chlorine and attacking oxygen of alkoxide. The latter mechanism omits carbonyl oxygen which is an integral part of the reaction and responsible for the reactivity of the reacting molecule.

To overcome these limitations regarding the intrinsically approximate answer (ELECTRO) and the scalability (Reaction Explorer) of existing methods, we here propose a two-step method called *MechFinder* to label reaction mechanisms in a chemically reasonable and scalable manner using automatically extracted reaction templates and expert-coded mechanistic templates. In our method, we extract the most important subgraph of reaction transformation in a reaction template, and manually label the mechanism of each reaction template based on chemist knowledge in the form of mechanistic template. This adoption of dual templates (reaction templates and mechanistic templates) allows us then to label the mechanisms of a large number of chemical reactions automatically by going through the two separate models to determine the types of templates. Our method enables the generation of mechanistic pathways for various reaction types, including pericyclic reactions and those involving multiple valence bond changes, such as reductive amination. These labels serve as computational analogs to the conventional arrows employed in the arrow-pushing model. To evaluate the effectiveness of our method, we curated a subset of the USPTO-50K dataset as a benchmark and assessed the coverage and applicability of our approach.

The main contribution of our work is three-fold:

1. We proposed *mechanistic template*, a complementary template for human chemist to classify and encode the reaction mechanism for each arbitrary *reaction template*.
2. Based on the reaction templates and mechanistic templates, we present *MechFinder*, the first scalable rule-based model to automatically label the chemically reasonable reaction

mechanisms. The generated reaction mechanisms by *MechFinder* are shown to be chemically much more reasonable than those of the previous method.

3. We curate a high-quality reaction dataset with chemically reasonable reaction mechanisms, denoted here as mech-USPTO-31k, which would benefit the chemistry community to develop prediction models for the reactivities that are based on molecular mechanisms.

2 Methods

2.1 Reaction Template (RT)

In our approach, we leverage the insight that many seemingly different organic reactions often follow similar patterns of electron flow, known as reaction rules, which are localized around specific atoms and bonds. This allows us to narrow down the scope of deriving mechanistic labels by focusing only on the atoms involved in the reaction. To obtain the reactivity information of a reaction dataset, we extract a set of reaction templates (RT) from each reaction in the dataset based on the local reaction template proposed in LocalRetro [15] based on RDChiral [16]. We start by identifying the reaction centers by comparing the chemical environments between the same atoms before and after the reaction. Nonetheless, we recognize that in many cases the electron movement can go beyond the changed atoms, such as the nucleophilic acyl substitution reaction shown in Figure 1. Therefore, we also include moieties that are π-conjugated to the changing atoms, such as double, triple, and aromatic bonds, and several mechanistically important special groups, such as carbonyl group and acetal group. This reaction template is simpler than the template extracted by RDChiral [16] but more informative than the local reaction template described in LocalRetro [15]. The overall template extraction is performed by the following five steps:

1. Compare the chemical environment of each atom before and after the reaction according to the atom-mapping (reactant-product atom correspondence). The atoms found to have changes in chemical environment are identified as "changed atoms".
2. For each identified changed atom, we identify the neighboring atoms connected to the changed atom in the reactants with double, triple, or aromatic bond as "extended atoms".
3. To further extend the scope of RT for mechanism labeling, we manually define a set of mechanistically important special groups. If any of the changed atoms are identified in one of the special groups, all the atoms in the special groups are also added to the "extended atoms" list. The RT extension process is illustrated in Figure S1. The set of defined special groups can be found in Figure S2.
4. After identifying the extended atoms in the reactants, we record the atoms sharing the same atom-map numbers in the product.
5. Using RDKit python package [17], we extract the chemical fragment in the reactants and products in SMARTS format based on the identified changed atoms and extended atoms, and connect the fragments by a reaction symbol ">>".

The full list of top-100 RTs can be found in Table S1.

2.2 Mechanistic template (MT)

Since RTs only capture the changes before and after the reaction, simply applying heuristic rules on RTs to generate mechanistic pathway without any in-domain chemistry knowledge poses clear limitations, as the example shown in Figure 1. Therefore, we additionally introduce the concepts of mechanistic classes (MC) and mechanistic templates (MT) to describe the actual reaction mechanism. The MC is defined as a group of reactions following the same reaction mechanism, including one or multiple RTs. For a given MC, we then hand-code the MT which describes the direction of electron movements in the form of a sequence of arrow-pushing diagrams, representing the attacking and electron-receiving moieties to incorporate chemistry knowledge.

The proposed MTs are represented by categorizing the arrows that illustrate the movement of electron pairs in organic reactions into four groups: lone pair to atom, lone pair to bond, bond to atom, and bond to bond. Technically, the lone pairs of atoms are simply annotated by their atom-map numbers

and the electron pairs from bonds are annotated by pairs of atom-map number. The full list of 63 MTs hand-coded from top-100 RTs extracted from the USPTO dataset can be found in Table S2.

The proposed MT has four notable features: (1) Because the atom types are specified in RTs but not in MT, multiple RTs often share the same MT. For example, different nucleophiles in substitution reactions can lead to different RTs but the same MTs (Figure 2a). An example illustrating this feature is provided in Figure S3. (2) In some cases, a single RT can match different MTs depending on the specific chemical environment. In these cases, we design particular criteria to assign the correct MT to the obtained RT. For example, the decision of assigning S_N1 and S_N2 depends on the alkane group connected to the leaving group (Figure 2b). List of all criteria used in *MechFinder* along with example reactions can be found in Figure S4. (3) For many reactions, the reaction can only occur when additional reagents are added, and the reaction mechanism can only be labeled if these reagents exist. For these reactions, we put the necessary additional reagents into the reactant set to complete the mechanism (Figure 2c). (4) Since the mechanistic pathway labeled by this method is based on the movement of electron pairs, reaction mechanisms beyond this scope such as organometallic or radical reactions cannot be labeled by the current method (Figure 2d).

It is noteworthy that, the mechanism derivation for certain groups of reactions inevitably requires the involvement of additional moieties beyond those present in the extracted RT. To address the limitation associated with the locality, we have incorporated technical maneuverability into our method to capture the important mechanistic elements. The framework and examples can be found in Figure S5.

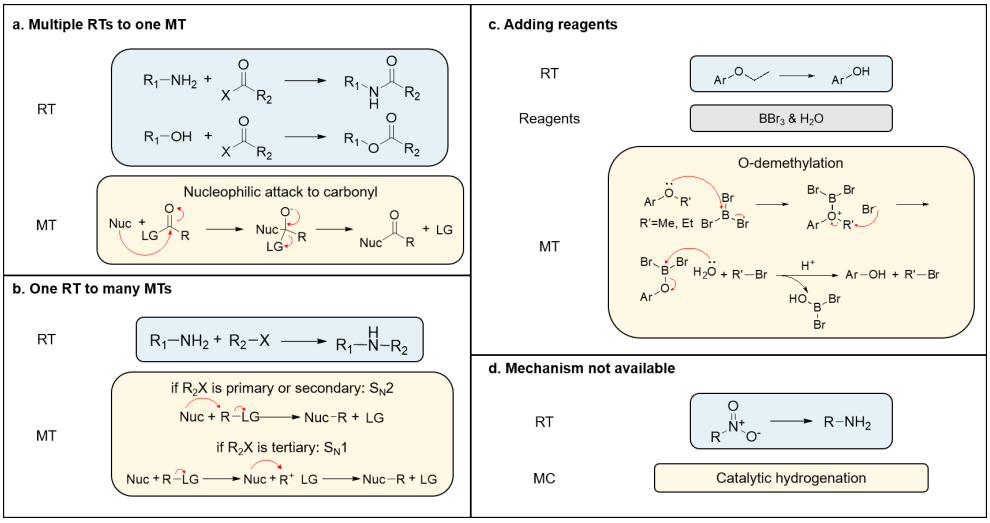

Figure 2: The four features of a proposed mechanistic template (MT). (a) Multiple RTs can match a single MT if they share the same mechanism. (b) One RT can lead to different MTs according to the criteria of the belonging MT. (c) Necessary reagents are added to complete the mechanistic pathway labeling. (d) Reactions whose mechanisms do not follow an arrow-pushing diagram cannot be labeled.

3 Results

3.1 Mechanistic annotation based on dual templates

In this paper, we introduce a mechanism labeling framework called *MechFinder* utilizing RTs and MTs introduced above. The process of using *MechFinder* to label the reaction mechanisms in a reaction dataset is divided into two phases: the expert annotation phase and the automatic labeling phase, as shown in Figure 3a and 3b.

During the expert annotation phase (Figure 3a), we first extracted N (N = 100 in the current dataset used) unique RTs from all the X reactions (X = 33,099 in the current dataset used) in the reaction

dataset. For each RT, we sampled k representative reactions to manually label the mechanism by three steps shown in Figure 3c:

1. **RT extraction**. We extracted reaction template focused on the reaction center, describing the local changes in atomic configuration upon a chemical transformation. The extraction process also yields an atom-map lookup table, recording the one-to-one atom-map correspondence between the input reaction and the extracted RT.
2. **MT identification**. Having RT for the given reaction, the MC and MT is identified by manual labeling in the expert annotation phase (but, once mapped, automated in the actual large-scale mechanism generation).
3. **Mechanistic sequence acquisition**. The mechanistic pathway for the input reaction is labeled by aligning the atom-map numbers from the MT to the input reaction according to the atom-map lookup table.

The number of sampled reactions k in the expert annotation phase is defined by the complexity of the encountered RT. For simple reactions like nucleophilic acyl substitution, we only sample one reaction to label the MT. For more complex reactions like S_NAr reaction, we sample more reactions to include more cases where the electron withdrawing groups (EWG) are located at different positions (ortho or para) with respect to the leaving group to label the MT with different criteria.

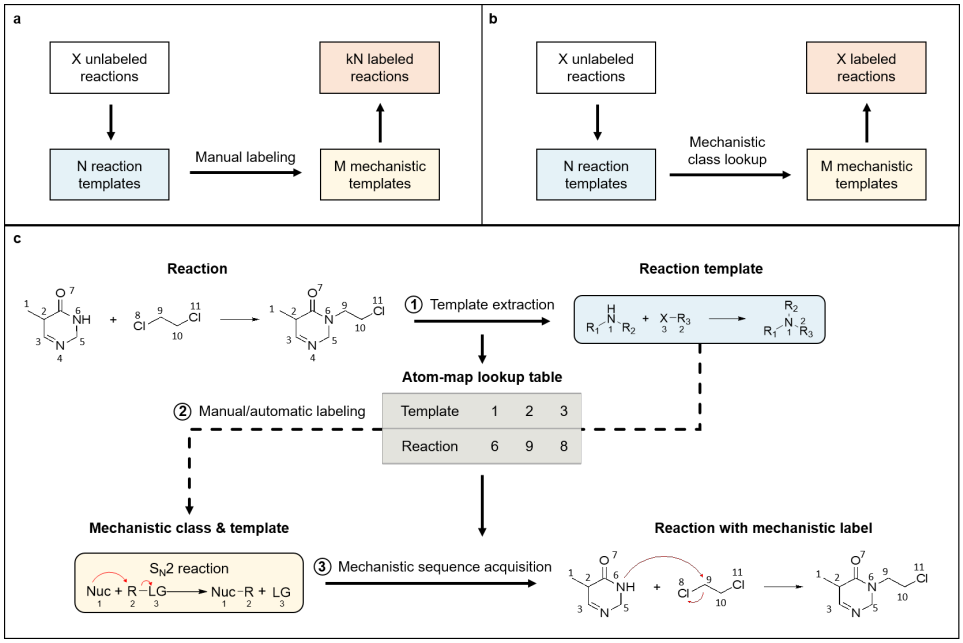

Figure 3: Overall pipeline of (a) expert annotation and (b) automatic mechanism labeling with *MechFinder*. (c) Example of the process of labeling a S_N2 reaction. First, we extract the RT of the reaction focused on reacting atoms (nitrogen, alkyl group, and chlorine with atom-map numbers of 6, 9, and 8, respectively) along with the lookup table containing one-to-one correspondence between template map numbers and atom-map numbers in the original reaction. Next, we manually identify the MC for extracted LRT as "S_N2 reaction" which has its unique MT characterized by the inherent electron flow of reactive moieties. Upon replacement of template map numbers in the MT (1, 2, 3) by the corresponding atom-map numbers in the original reaction (6, 9, 8) using the lookup table, we finally obtain mechanistic sequence for the given reaction.

During the automatic labeling phase (Figure 3b), we follow the same three steps described in the set-up phase (Figure 3c) but replacing the manual labeling step (step 2) by looking up the previously labeled MT for the identified RT during the manual labeling phase to label the reactions in the reaction dataset. More examples of automatic mechanism labeling by *MechFinder* can be found in Supplementary Information.

3.2 Dataset

In our benchmark experiments, we used the reaction data extracted from USPTO grant patents collected by Lowe [18], an organic reaction dataset extensively used in benchmarking various reaction prediction approaches. In particular, we demonstrate the results using USPTO-50K dataset curated by Schneider et al. [19] without reagent information. Since our approach only addresses arrow-pushing diagram representable mechanisms, we removed organometallic and radical reactions based on extracted RT. In addition, we applied LocalMapper to refine the atom-mapping of reactions because *MechFinder* is highly dependent on the quality of correct atom-mapping. Thus, 33,099 reactions are finally obtained after the above pre-processing procedure. We refer to this reaction subset as USPTO-33K dataset in this paper.

3.3 Quantitative results

From a total of 33,099 reactions in the USPTO-33K dataset, we identified 400 RTs based on their precise atom-mappings. Among these, we categorized the 100 most frequently occurring RTs (shared by the most chemical reactions) into 63 distinct MCs (X=33,099, N=100, M=63) aligned with the conceptual framework of arrow-pushing-diagram-representable reaction mechanisms. The chosen top 100 RTs cover 94.8% of the USPTO-33K dataset reactions. To this end, our method generated the mechanistic annotation of 31,364 reactions that can be described by arrow-pushing-diagrams. This is the first large-scale mechanism dataset for organic reactions in the present literature, and we denote it as mech-USPTO-31k.

We analyze the number of labeled reactions by increasing the number of labeled MTs and labeled RTs between top-1 to top-100 most popular RTs in Figure 4a. We found that labeling the top 10 most popular RTs with 11 MTs can successfully cover 58.4% of the total reactions, and labeling the top 50 most popular RTs with 40 MTs can cover 87.7% of the total reactions. The coverage increments of adding a new MT drop exponentially with the decreasing RTs popularity. We inspect the frequency of the reactions in the mech-USPTO-31K dataset relative to the obtained label length to gain insight toward the complexity of the labeled reactions in terms of the lengths of the mechanistic pathways (Figure 4b). The majority of reactions exhibit 2-, 4-, 8- and 12-steps, which are mainly within the top-10 RTs. Remarkably, certain reactions feature lengthy sequences such as Swern oxidation [20] in 19 steps, multi-component imidazole synthesis from a carboxylic acid and diamine substrate in 22 steps and Vilsmeier formylation [21] in 23 steps. These findings underscore the versatility of our labeling method, which accommodates a wide spectrum of reactions, irrespective of their intricacy.

The top 10 most popular RTs and their corresponding MTs are shown in Figure 4c. The most popular RT, covering 16.6% of the reactions in the mech-USPTO-31k dataset, corresponds to DCC condensation, which requires 12 mechanistic steps. The top-2 and top-8 RTs indicate S_NAr reactions but use different nucleophiles. Similarly, the top-3 and top-5 RTs indicate S_N1 or S_N2 reactions. The mechanisms of top-4 and top-10 RTs can be represented by the MTs showing nucleophilic attack to the carbonyl (or sulfonyl) group. The top-6 is Boc group deprotection in acid and the top-7 RT is reductive amination.

4 Discussion and Limitations

Since we only consider organic reactions, whose mechanisms can be represented by arrow-pushing diagrams showing the movement of electron pairs, the current capacity of the proposed method is limited by the inability to label the mechanisms of organometallic and radical reactions. In the case of organometallic reactions, proposed mechanisms are alternate steps of one or more among oxidative addition, migratory insertion, reductive elimination, and β-hydride elimination which follow different patterns of electron movements. Mechanistic labeling of those reactions should include a model with different computational representations of electron movements. Building such model reflecting those patterns would enable extending the coverage to include organometallic reactions in the labeled dataset. The case of radical reactions is similar in the sense that, some modification to our current mechanistic representation to incorporate fishhook arrows showing the movement of single electrons would enable their labeling as well.

Although not addressed in this article, our method should be further improved to include stereochemical analysis. The popularity and importance of stereo- and enantio-selective syntheses in chemical

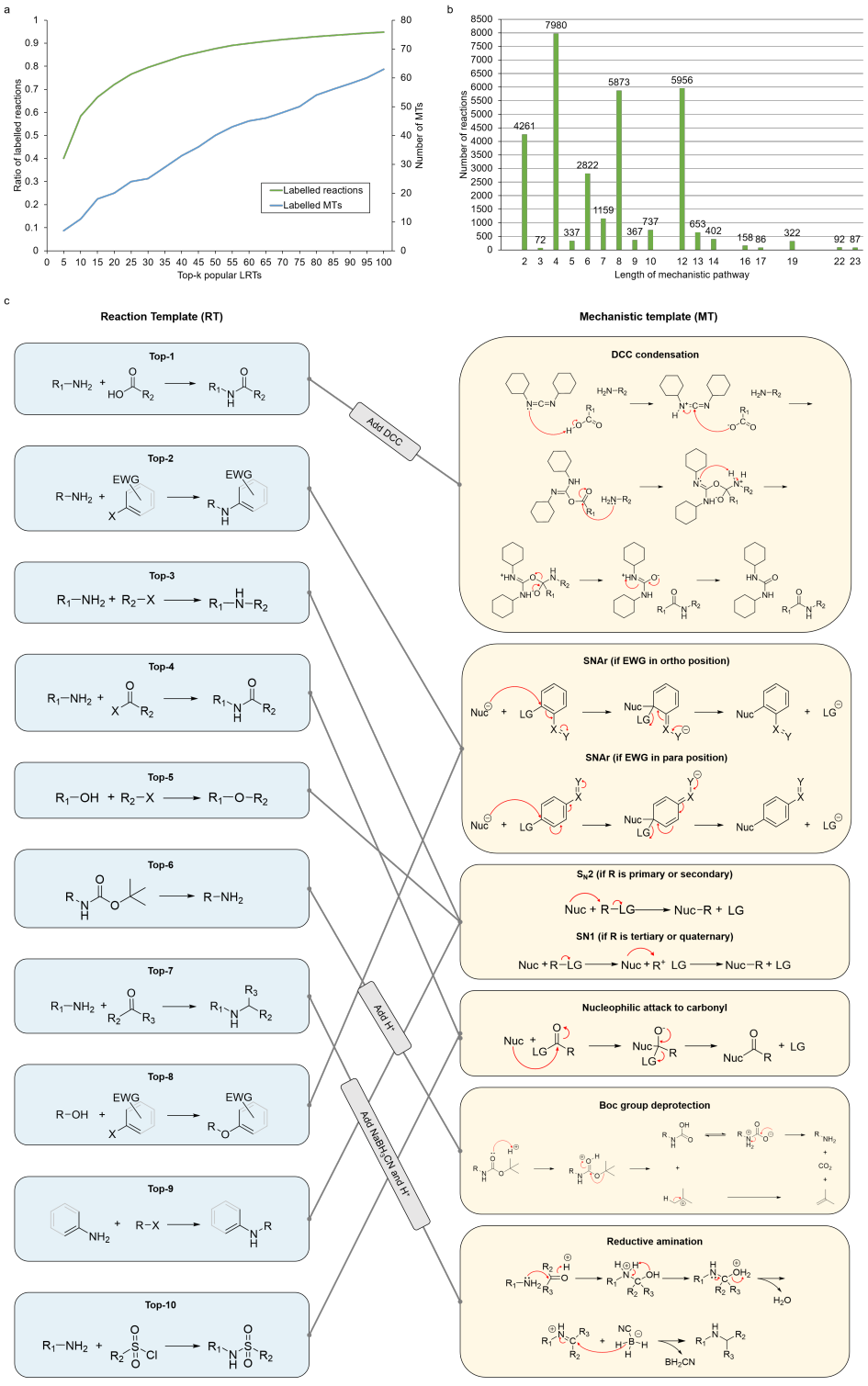

Figure 4: The quantitative results of automatic generation of mechanistic pathways for the USPTO-33K reaction dataset using *MechFinder*. (a) Number of labeled reactions and MTs as a function of the number of most popular RTs from 1 to 100. (b) Distribution of mechanistic pathway lengths, displaying the frequency of n-step reactions in the labeled dataset. (c) The top 10 most popular RTs and the 6 MTs associated with each RT. Note that RTs are expressed in SMARTS format and do not contain any hydrogen. The depictions are simplified for reading convenience.

and drug industry make comprehensive stereochemical elucidation necessary [22]. Because the stereochemical outcome of an organic reaction is governed by the stereochemistry of each mechanistic step, one may account for reaction stereochemistry by extending arrow-pushing diagrams with stereochemical definitions [23]. In that way, the overall stereochemistry of the reaction can be explicated.

5 Conclusion

In this work, we presented a computational approach, *MechFinder*, to automatically generate reasonable mechanistic sequences for organic reactions by capturing the patterns of electron movements in sequence from the large reaction dataset. An expert-coded mechanistic dictionary containing a many-to-one mapping of extracted reaction templates (RTs) and mechanistic templates (MTs) provides the basis for the accurate acquisition of so-called mechanistic labels, or reaction mechanisms. With the direct utilization of this technique we were able to reliably generate reaction mechanisms for nearly 95% of the USPTO-33k dataset automatically by manually labeling 100 RTs and 63 MTs. This is the first systematic mechanism dataset of this scale for organic reactions which we denote as mech-USPTO-31k. The reaction coverage of the method can be further improved by defining computational representations for organometallic and radical reactions. Having a large-scale mechanism dataset such as mech-USPTO-31k automatically generated here, we envisage that an interpretable and more reliable machine learning model for reaction prediction can be built and trained on it. Since these mechanism-based ML models will make predictions based on the plausible mechanisms, rather than based on the learning just the reactant and product information, they will have less bias towards the known reactions and potentially provide an exciting opportunity to develop and discover new chemical reactions beyond human intuitions.

References

[1] A Filipa de Almeida, Rui Moreira, and Tiago Rodrigues. Synthetic organic chemistry driven by artificial intelligence. *Nature Reviews Chemistry*, 3(10):589–604, 2019.

[2] Wendy A Warr. A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. *Molecular Informatics*, 33(6-7):469–476, 2014.

[3] Stuart Warren and Paul Wyatt. *Organic synthesis: the disconnection approach*. John Wiley & Sons, 2008.

[4] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[5] Shuan Chen and Yousung Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022.

[6] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

[7] Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.

[8] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575, 2020.

[9] Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

[10] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

[11] Venkat Venkatasubramanian and Vipul Mann. Artificial intelligence in reaction prediction and chemical synthesis. *Current Opinion in Chemical Engineering*, 36:100749, 2022.

[12] Daniel E Levy. *Arrow-pushing in organic chemistry: an easy approach to understanding reaction mechanisms*. John Wiley & Sons, 2017.

[13] Jonathan H Chen and Pierre Baldi. No electron left behind: a rule-based expert system to predict chemical reactions and reaction mechanisms. *Journal of chemical information and modeling*, 49(9):2034–2043, 2009.

[14] John Bradshaw, Matt J Kusner, Brooks Paige, Marwin HS Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. *arXiv preprint arXiv:1805.10970*, 2018.

[15] Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.

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

[17] The RDKit Development Team. *The RDKit Documentation*, 2020. Accessed: [2023].

[18] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[19] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

[20] Kanji Omura and Daniel Swern. Oxidation of alcohols by "activated" dimethyl sulfoxide. a preparative, steric and mechanistic study. *Tetrahedron*, 34(11):1651–1660, 1978.

[21] Siegfried R Waldvogel. Comprehensive organic name reactions and reagents. *Synthesis*, 2010(05):892–892, 2010.

[22] Hans-Jürgen Federsel. Asymmetry on large scale: the roadmap to stereoselective processes. *Nature Reviews Drug Discovery*, 4(8):685–697, 2005.

[23] Kurt Mislow. Stereochemical terminology and its discontents. *Chirality*, 14(2-3):126–134, 2002.