CompRet: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration

Ryosuke Shibukawa,†△ Shoichi Ishida,‡△ Kazuki Yoshizoe,§ Kunihiro Wasa,§
Kiyosei Takasu,‡ Yasushi Okuno,∥⊥ Kei Terayama,*,#,¶,∥,⊥ and Koji Tsuda,*,†,#,¶

†Graduate School of Frontier Sciences, the University of Tokyo, Kashiwa, 277-8561, Japan
‡Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, 606-8501, Japan
¶RIKEN Center for Advanced Intelligence Project, Tokyo, 103-0027, Japan
§Toyohashi University of Technology, Toyohashi, 441-8580, Japan
∥Graduate School of Medicine, Kyoto University, Sakyo-ku, 606-8507, Japan
⊥Medical Sciences Innovation Hub Program, RIKEN, Yokohama, 230-0045, Japan
#Graduate School of Medical Life Science, Yokohama City University, Tsurumi-ku,
230-0045, Japan
@Research and Services Division of Materials Data and Integrated System, National
Institute for Materials Science, Tsukuba, 305-0047, Japan

△These two authors contributed equally

E-mail: terayama@yokohama-cu.ac.jp; tsuda@k.u-tokyo.ac.jp
Phone: +81 (0)4 71363983. Fax: +81 (0)4 71363982

Abstract

In computer-assisted synthesis planning (CASP) programs, providing chemical synthetic routes as many as possible is essential for considering optimal and alternative
routes in a chemical reaction network. As the majority of CASP programs have been
designed to provide one or a few optimal routes, it is likely that desired one will not
be included. To avoid this, an exact algorithm that lists possible synthetic routes from
the chemical reaction network is required, alongside a recommendation of synthetic
routes that meet specified criteria based on chemist’s objectives. Herein, we propose
a chemical-reaction-network-based synthetic route recommendation framework called
“CompRet” with a mathematically guaranteed enumeration algorithm. In a prelim-
inary experiment, CompRet was shown to successfully provide alternative routes for
a known antihistaminic drug, cetirizine. CompRet is expected to promote desirable
enumeration-based chemical synthesis searches and aid the development of an interac-
tive CASP framework for chemists.

1 Introduction

Since the 1960s, several researchers have proposed computer-assisted chemical synthetic
route designs. Various computer-assisted synthesis planning (CASP) programs have been
developed to assist synthetic organic chemists in their work.\textsuperscript{1–3} While expert systems and
knowledge-based programs were primary focus of CASP during the early stages,\textsuperscript{4–8} recent
breakthroughs in the field of deep learning have accelerated its development.\textsuperscript{9–17} In particular,
data-driven approaches have received attention across research fields.\textsuperscript{18,19} These approaches
have shown outstanding performances at every stage, and more recently, they have provided
realistic and preferable synthetic routes.

The pioneers of CASP, Corey and Wipke, stated the following requirements related to
the above strategy in their paper:\textsuperscript{2} the program needs to provide as many useful routes as
possible, chemists can decide the depth of search or analysis of the synthetic route, and
the numerous routes are evaluated by the chemists. As discussed above, several CASP
approaches have been developed; however, the majority of them have aimed to directly
obtain the optimal chemical synthetic route rather than attempting to provide multiple route
candidates. According to Corey, examining chemical synthetic routes as many as possible is an essential part of retrosynthetic analysis. It is well known that evaluation criteria on presented synthetic routes depend on the chemist’s situation, objectives, and/or needs, such as the early-stage derivatization of hits, optimization of lead compounds, or large-scale synthesis of drug candidates. Thus, a desirable framework should provide as many routes as possible under specific conditions and choose multiple reliable routes based on the given situation.

As a framework for providing multiple reliable routes, Kowalik et al. have developed a promising approach using the Network of Organic Chemistry (NOC) and an enumeration algorithm of possible synthetic routes. The NOC consists of all possible molecules and reactions that represent links from reactants to products. The reactions are practically represented as templates that include the conditional/contextual rules of chemistry. Additionally, they realized a recommendation system of multiple synthetic routes for a target molecule as follows. Firstly, they extracted the network of molecules and reactions (chemical reaction network) related to the target from the NOC. Secondly, they enumerated all possible synthetic routes from the chemical reaction network and then selected promising candidate routes. Although they showed a huge number of synthetic routes for some molecules and presented realistic solutions, their approach has two potential issues: the NOC is very large and is thus uneconomical for obtaining optimal routes for a specific target molecule, and the enumeration algorithm does not always exactly provide all possible routes. Hence, an efficient algorithm for constructing a chemical reaction network is required for practical application. Further, an exact enumeration algorithm without loss and duplication is needed for practical usage and finding reliable alternative routes.

In this study, we propose a CASP framework called “CompRet,” which provides multiple possible routes using a novel enumeration algorithm with a theoretical guarantee. CompRet implements the following three steps to recommend synthetic routes: (1) constructing a chemical reaction network based on the depth-first proof number search (DFPN) and
template-based retrosynthesis,\textsuperscript{27,28} without a large chemical reaction network such as the NOC, (2) enumerating all synthetic routes from the network using a novel algorithm, and (3) recommending multiple synthetic routes by developing a naive visualization method and simple scoring functions. DFPN was initially developed by Nagai et al. in the context of artificial intelligence for games such as Shogi and Go.\textsuperscript{29,30} On application to CASP, it has shown superior or comparable performance to that of depth-first search or Monte-Carlo tree search.\textsuperscript{31,32} Therefore, DFPN was adopted to construct the chemical reaction network proposed herein. As a result of the enumeration, the number of possible synthetic routes may reach or exceed several millions. As it would be impossible for chemists to manually examine all of them, several scoring functions and a visualization method have been introduced into the framework to simplify the process.

Here, we report the development of CompRet and mathematical prove that the proposed enumeration algorithm, which can exactly enumerate all synthetic routes from a given constructed chemical reaction network. To demonstrate the approach, possible synthetic routes were found for cetirizine, an antihistaminic drug. In addition to sorting routes by scores, an embedding method to obtain an overview of millions of synthetic routes by defining a route fingerprint was attempted.

2 Methods

2.1 Construction of chemical reaction network

A synthetic route for a target molecule can be represented as a tree-like structure in which molecule nodes (circles) and chemical reactions (rectangles) appear alternately, as shown in Fig. 1 (a). In order to make a route feasible, the end molecule nodes (molecules in the blue circles in Fig. 1 (a)) must consist of starting materials (e.g., commercially available molecules), and each reaction step should be reasonable.\textsuperscript{27,28}

The chemical reaction network of a target molecule is typically large and can efficiently
Figure 1: AND/OR representation of synthetic routes. (a) Example of a single synthetic route for cetirizine. Retrosynthesis computation is performed recursively from the topmost node (target) until it reaches to starting materials at the bottom. (b) The synthetic route can be represented as an AND/OR tree in which the OR and AND nodes denote molecules and reactions, respectively. (c) A chemical reaction network ideally consists of all possible synthetic routes of a target (red sphere) molecule.

express (ideally all) possible synthetic routes to the target represented by the molecule and reaction nodes. In this study, we represent synthetic routes and chemical reaction networks as AND/OR trees, as shown in Fig. 1 (b), to efficiently construct chemical reaction networks and exactly perform enumeration (see Supporting Information for details on the AND/OR tree and chemical reaction network). The synthetic route in Fig. 1 (a) is represented as the black route in Fig. 1 (b) by expressing the molecule and reaction nodes as OR and AND nodes, respectively. The gray route in Fig. 1 (b) shows another route to synthesize the same reactant. In an AND/OR tree, a molecule is represented as an OR node because the molecule can be synthesized from the either black “OR” gray routes, as shown in Fig. 1 (b). On the other hand, a reaction is represented as an AND node because all the reactants (e.g.,
B “AND” C OR nodes in Fig. 1 (b)) of the reaction (AND node A in Fig. 1 (b)) are required to synthesize the product. Merging the molecule and reaction nodes that appear in different synthetic routes in this manner (Fig. 1 (b)) enables an efficient representation of a large number of routes as a chemical reaction network (Fig. 1 (c)).

CompRet efficiently constructs the chemical reaction network for a given target molecule based on DFPN, a search method based on the AND/OR tree using proof and disproof numbers on the nodes (see Supporting Information for details on the DFPN algorithm). To design a synthetic route, reaction templates are applied to a target to transform it into reactants. For the retrosynthetic computation, Reactor version 20.11.0 (ChemAxon) was used. The relevance of the transformed reactants was checked by computing the product of the template and the reactants. By recursively performing this transformation according to the DFPN algorithm with the preset maximum depth \( md \), a possible synthetic route for a target molecule can be obtained. The algorithm can design longer synthetic routes with a larger \( md \) value. Furthermore, CompRet repeatedly searches for a new route and merges it into a chemical reaction network (see Supporting Information for details on the construction algorithm).

### 2.2 Enumeration algorithm

Enumerating all synthetic routes in the chemical reaction network of a given target may appear to be a simple problem, as described in the literature reported by Kowalik et al.\(^{26}\) For example, in Fig. 2 (a), the target (molecule 1) can be synthesized via one of the reactions A, B, or C. Here, we consider how to count all possible synthetic routes. Synthetic routes for the target can be counted using the following equation:

\[
mol(1).count = rxn(A).count + rxn(B).count + rxn(C).count,
\]  

(1)
where \( \text{mol}(1).\text{count} \) and \( \text{rxn}(X).\text{count} \) denote the number of routes to synthesize the molecule 1 and the number of routes that use the reaction \( X \), respectively. On the other hand, the number of ways to prepare reactants for a reaction is calculated as follows. In the case of \( A \),

\[
\text{rxn}(A).\text{count} = \text{mol}(2).\text{count} \times \text{mol}(3).\text{count}. 
\]

These calculations are performed recursively until starting materials for which \( \text{mol}(\cdot).\text{count} \) is assigned one (e.g., \( \text{mol}(3).\text{count} = 1 \)).

However, as mentioned in the literature, this procedure does not count the exact number of synthetic routes because it assumes that each reactant is synthesized independently. In the actual network, a single molecule can act as a reactant for several reactions; for example, the molecule 4 in Fig. 2 (b). Note that both molecules 2 and 3 are required for the reaction \( A \). The network depicted in Fig. 2 (b) includes only two synthetic routes for the molecule 1, i.e., a choice of the reactions \( D \) or \( E \) for synthesizing the molecule 4, while the number of synthetic routes is calculated to be 4 according to the above equations. Therefore, we propose an enumeration algorithm that extracts all possible routes in a network by considering “joined nodes”, as shown in Fig. 2 (b). The details of the enumeration algorithm are described in the Supporting Information.

### 2.3 Route ranking for recommendation

Three scoring functions were utilized for recommendation from the enumerated synthetic routes: the step-based method (STEP), mean synthetic complexity score (MSCS), and reference route-based method (REF). STEP is a simple method that outputs the largest number of reaction steps for a given synthetic route. Synthetic complexity score (SCScore) was developed by Coley et al.\(^{12}\) to evaluate the complexity of the molecule. Like the SCScore, the MSCS for a synthetic route ranges from 0 to 5. As the SCScore of a molecule is directly proportional to the complexity of its synthesization, a molecule with a lower SCScore is
Figure 2: Example of local structure of chemical reaction network to explain the method of calculating the number of synthetic routes. (a) Ideal structure of a network for which the naive method can count the exact number of synthetic routes for a target molecule. (b) The naive method cannot count synthetic routes exactly in this case. The number of synthetic routes for the target molecule, \( \text{mol}(1).\text{count} \), is calculated as \( \text{mol}(5).\text{count} = 1, \text{mol}(6).\text{count} = 1, \text{rxn}(D).\text{count} = \text{mol}(5).\text{count} = 1, \text{rxn}(E).\text{count} = \text{mol}(6).\text{count} = 1, \text{mol}(4).\text{count} = \text{rxn}(D).\text{count} + \text{rxn}(E).\text{count} = 2 \), and finally \( \text{mol}(a).\text{count} = 4 \), while the true number of synthetic routes is 2.

preferred in synthesis planning. Here, MSCS is defined as the average of the SC Scores of all molecules in a synthetic route. REF is calculated only if a reference synthetic route is given. First, all the molecules that appear in the reference route are extracted and sorted by molecular weight. A list of sorted molecules is similarly prepared for a designed synthetic route. Then, the sum of fingerprint-based similarities between the sorted molecules of the reference and given routes is calculated. The RDKit fingerprint\(^{35,36}\) and Tanimoto metrics\(^{37}\) were employed for this similarity evaluation. If the lengths of the sorted molecule lists differ,
REF is set to 0. Smaller STEP and MSCS values indicate a superior route, whereas REF is designed such that a larger value indicates a more desired route.

2.4 Visualization for confirming route distribution

To confirm that the CompRet framework is capable of designing a wide variety of synthetic routes, we developed a simple method to plot them on 2D space by converting a synthetic route into a vector. For the conversion, route fingerprint \( f_r \) for a route \( r \) is defined as

\[
f_r = \sum_{t \in r} f_p(t),
\]

where \( t \) is a reaction template and \( f_p(t) \) is the structural reaction fingerprint\(^{38} \) of \( t \) computed by RDKit.\(^{36} \) Following the computation of route fingerprints for 3,000 sampled routes, t-SNE embedding\(^{39} \) was computed using scikit-learn.\(^{40} \)

2.5 Reaction template and building block

Template-based approaches generally require both reaction templates and starting materials. Reaction templates are represented as a generalized chemical reaction and technically represented as a reactive center and the first neighboring atoms and bonds in a reaction. Reaction templates were extracted from 27 million single-step reactions obtained from Reaxys (from 1795 to 2019),\(^{41} \) following the method used in a previous study.\(^{17} \) Here, the single-step reactions obtained from Reaxys were filtered on the condition that a reaction has a product and up to three reactants. Five hundred reaction templates were used in the order of occurrence frequency. Starting materials are defined as commercially available chemical compounds and used as stopping criteria for DFPN. For these, 157,544 molecules from Enamine building blocks\(^{42} \) were used.
3 Results and Discussion

3.1 Proof of enumeration algorithm

The proposed algorithm can enumerate all possible routes without loss or duplication from a given chemical reaction network of a target molecule. To prove this, it is necessary to show that (1) the algorithm outputs only synthetic routes, (2) there are no duplicate outputs, and (3) the algorithm outputs all synthetic routes in a given chemical reaction network. Here, these properties have been proven using the partition method,\textsuperscript{43,44} which is widely used for enumeration algorithms, and mathematical induction. Details of these proofs are given in the Supporting Information. The properties (1), (2), and (3) are shown in Theorem 4.1, Lemma 4.1, and Theorem 4.2, respectively. The algorithm described in the literature reported by Kowalik et al.\textsuperscript{26} cannot count the number of synthetic routes accurately, as discussed in the Methods section. On the other hand, the proposed enumeration algorithm exactly outputs all synthetic routes without loss or duplication based on the idea of the “prohibited list” (variable $P$) and related procedures in Algorithm S3.

3.2 Route enumeration for Heifets’ benchmark

For the first demonstration, Heifets’ benchmark molecules\textsuperscript{31} were used to show the scale of the chemical reaction networks constructed by CompRet and the synthetic route enumeration for the networks. The top 100 reaction templates of the prepared template data were used, and $md$ was set to six. All calculations were conducted using a single CPU core (Intel(R) Xeon(R) CPU E5-2690 v3 @ 2.60GHz) with 256 GB of RAM.

Table 1 shows the results of the construction of chemical reaction networks and enumerations for the target molecules. Only those molecules whose synthetic routes were found in the top 100 templates among the Heifets’ benchmark. Note that the template set prepared in their study\textsuperscript{31} consisted of 50 reactions that were selected to be suitable for synthesis of the benchmark molecules, although they succeeded in finding synthetic routes for most
molecules in the benchmark. In Table 1, the second and third columns from the left indicate the number of constituent OR and AND nodes, respectively, that is, the size of the chemical reaction network for the molecule depicted in the first column. The sizes of the generated chemical reaction networks differed significantly between the molecules. The fourth column from the left shows the number of enumeration results of the synthetic routes extracted from each chemical reaction network. More than 100,000 synthetic routes have been successfully enumerated for the fifth and sixth molecules. This number may appear excessive considering the molecules; however, as reported in prior studies,\textsuperscript{25,26} the number of synthetic routes can reach $\approx 10^5$ depending on molecules. Thus, the obtained results are consistent with previous findings. The fifth and sixth columns denote the memory sizes of the constructed network and enumerated routes, respectively. Each object is converted into DOT format\textsuperscript{45} to calculate the total amount of memory. The seventh and eighth columns show the generation time of the chemical reaction network and the calculation time of enumeration, respectively. It can be seen that the time for enumeration tends to be much longer than that for the construction of the chemical reaction network. This is because searching for applicable templates for a molecule and then using them to divide it to its substances are time-consuming tasks. Besides, enumeration from a larger and more complex network tends to require more time because the number of synthetic routes in a chemical reaction network increases combinatorially.

| Molecule | OR Nodes | AND Nodes | Enumeration Results | Network Memory | Route Memory | Generation Time | Calculation Time |
|----------|----------|-----------|---------------------|----------------|--------------|-----------------|-----------------|
| 1        |          |           |                     |                |              |                 |                 |
| 2        |          |           |                     |                |              |                 |                 |
| 3        |          |           |                     |                |              |                 |                 |
| 4        |          |           |                     |                |              |                 |                 |
| 5        |          |           |                     |                |              |                 |                 |
| 6        |          |           |                     |                |              |                 |                 |
3.3 Route recommendation for cetirizine

To examine the synthetic routes designed by CompRet in detail, we have applied CompRet to cetirizine, a drug whose reported synthetic route is relatively simple.\textsuperscript{46,47} Here, the results of changing the template set size and the maximum depth \(md\) are shown, followed by the routes recommended by CompRet using three scoring methods: REF, RSCS, and STEP.

First, the construction of the chemical reaction network and the synthetic routes for different template sizes (top 50, 100, and 500) were investigated. The value of \(md\) was fixed to six. Figure 3 (a) shows the time taken for the network constructions. The dotted line indicates that the first route has been found. Finding a single route for the top 100 (orange line) and 500 (green line) template cases required an extended period of time, because the number of candidate routes increased exponentially with the increase in the number of templates. The blue line shows the result for size 50. In this case, the network construction was completed in approximately 30 seconds. In the cases of size 100 and 500, the number of routes increased significantly, and thus, enumeration was halted when the number of routes exceeded 1,500,000. In the case of size 500, the time taken to find 1,500,000 routes was less than 2,000 seconds.

Figure 4 shows the scattered routes with the t-SNE embedding of the obtained synthetic routes, respectively. For ease of viewing, 1,000 sampled routes are shown for each template size. Three synthetic routes sampled from the distant plot in Fig. 4 have different respective starting materials and reactions. Additionally, the precursors of cetirizine in the middle and left routes have carboxylic acid, while the precursor in the reported route,\textsuperscript{47} shown as the black sphere route in Fig. 5, has carboxylate ester.

The results of constructing the chemical reaction network with the respective \(md\) values of 4, 6, and 8 are shown in Fig. 3 (b). The size of reaction templates was fixed to the top 100. The blue line in Fig. 3 (b) is the result for \(md = 4\), where the chemical reaction network was constructed in less than 10 seconds, and 853 routes were obtained. The orange line is identical to that in Fig. 3 (a), that is, the result of construction with \(md = 6\). Enumeration
Table 1: Experimental result of chemical reaction network construction and synthetic route enumeration for Heifets’ benchmark molecules. The numbers of OR and AND nodes indicate contained each node in the constructed chemical reaction network. The number of synthetic routes represents the enumerated number of possible routes in the network. The memory sizes of the network and total routes are calculated by converting objects into DOT files. The construction time indicates the computation time required for the construction of the network. The enumeration time indicates the computation time required to enumerate all possible synthetic routes from the network.

| Target molecule | The number of OR nodes | The number of AND nodes | The number of synthetic routes | Memory size of the network | Memory size of total routes | Construction time (sec) | Enumeration time (sec) |
|-----------------|------------------------|-------------------------|-------------------------------|----------------------------|---------------------------|------------------------|------------------------|
| ![Molecule Image](https://via.placeholder.com/150) | 29                     | 61                      | 70                           | 10.9 KB                    | 58.6 KB                   | 2.68                   | 2.91×10⁻²               |
| ![Molecule Image](https://via.placeholder.com/150) | 66                     | 153                     | 781                          | 27.5 KB                    | 1.15 MB                   | 6.02                   | 1.85×10⁻¹               |
| ![Molecule Image](https://via.placeholder.com/150) | 8                      | 11                      | 8                            | 2.11 KB                    | 7.33 KB                   | 2.96                   | 9.26×10⁻⁴               |
| ![Molecule Image](https://via.placeholder.com/150) | 22                     | 29                      | 32                           | 5.38 KB                    | 22.9 KB                   | 2.57                   | 9.01×10⁻³               |
| ![Molecule Image](https://via.placeholder.com/150) | 92                     | 376                     | 127,707                      | 62.6 KB                    | 311 MB                    | 11.7                   | 6.35×10¹                |
| ![Molecule Image](https://via.placeholder.com/150) | 34                     | 78                      | 108,385                      | 13.1 KB                    | 416 MB                    | 6.33                   | 9.76×10¹                |
| ![Molecule Image](https://via.placeholder.com/150) | 17                     | 19                      | 12                           | 4.00 KB                    | 6.77 KB                   | 4.30                   | 6.20×10⁻⁴               |
| ![Molecule Image](https://via.placeholder.com/150) | 38                     | 58                      | 67                           | 11.0 KB                    | 64.5 KB                   | 3.84                   | 1.46×10⁻¹               |
| ![Molecule Image](https://via.placeholder.com/150) | 74                     | 213                     | 5,213                        | 36.4 KB                    | 9.19 MB                   | 6.35                   | 6.45×10⁻¹               |
| ![Molecule Image](https://via.placeholder.com/150) | 45                     | 93                      | 1,042                        | 16.3 KB                    | 178 MB                    | 27.5                   | 3.93×10⁻¹               |
| ![Molecule Image](https://via.placeholder.com/150) | 65                     | 139                     | 2,520                        | 25.9 kB                    | 5.10 MB                   | 9.85                   | 3.46×10⁻¹               |
Figure 3: Searched synthetic routes for cetirizine are enumerated, and the number of the routes for each chemical reaction network is counted at several time points. (a) The labels Size 50, Size 100, and Size 500 respectively indicate that the top 50, 100, and 500 reaction templates have been utilized. (b) The value of $md$ shows the maximum depth of the search algorithm. Counting is aborted when the total number of routes exceeds 1,500,000. The dotted line at the bottom indicates that the first synthetic route has been discovered.

was halted when the number of routes exceeded 1,500,000. Figure S3 depicts the scattered routes explored with each $md$ using t-SNE embedding and several examples of them. This visualization also shows a variety of routes explored by CompRet.

Finally, the results of route recommendation using the route scoring methods REF, MSCS, and STEP are shown in Fig. 5. The figure illustrates several examples of synthetic routes recommended by these methods, explored using the top 500 templates and an $md$ of 6. CompRet successfully obtained a known efficient synthetic route,\textsuperscript{47} denoted by the black sphere in Fig. 5. Using this known route as reference and REF methods, CompRet recommended the yellow route that changed chloride to bromine in the reference route. Its similarity score was the highest (6.32) among the enumerated routes. This result indicates that the REF method is suitable for finding routes similar to the reference route. Using MSCS, CompRet recommended the orange and purple routes, whose MSCS scores were 1.80 (lowest) and 2.28, respectively. The two routes are different from the reference route; thus, MSCS method would have the potential to provide a variety of synthetic routes. Using STEP, CompRet recommended the red and light blue routes whose number of steps was smallest.
The green route also has a smaller number of reaction steps. While these routes had smaller numbers of reaction steps, their starting materials and reactions themselves differed. Consequently, the recommendation of routes that differed slightly from the known route as well as diverse alternatives suggests that the effectiveness of the CompRet framework.
Figure 4: Sampled route distribution by t-SNE embedding. Blue, orange, and green points denote sampled routes from the network constructed with the top 50, 100, 500 reaction templates, respectively. \( md \) is fixed to 6. For each setting, 1000 routes are sampled out of millions of candidates. The black rhombus indicates a chemically unreasonable reaction. Such a reaction may sometimes occur because CompRet algorithmically enumerates synthetic routes.
Figure 5: Examples of multiple synthetic routes for cetirizine recommended by CompRet. Each synthetic route is distinguished based on color. Parts of some routes may overlap, denoted by split spheres. In particular, the black sphere depicts the reported routes of cetirizine. 

---

17
4 Conclusions

In summary, we developed CompRet, a new recommendation framework for CASP. It consists of three parts: DFPN-based chemical reaction network construction of a given target, enumeration of synthetic routes from a given chemical reaction network, and recommendation from the enumerated synthetic routes. We have mathematically proven the validity of the enumeration algorithm. This algorithm works in general for a given appropriate chemical reaction network in general with a mathematical guarantee. CompRet was also applied to the Heifets’ benchmark molecules and cetirizine. It was demonstrated to be able to construct chemical reaction networks, containing over a million routes in some cases, with a relatively small computational cost. Furthermore, the recommendation and visualization methods could be useful to suggest a wide variety of conceivable synthetic routes, from a slight deviation from the existing route to significantly different alternatives.

CompRet is a fundamental framework; therefore, each module, particularly the construction of chemical reaction networks and recommendation methods by scoring, could be improved. As shown in the demonstration of chemical reaction network constructions, the explored synthetic routes depend on the template set and maximum depth $md$. The set of starting materials also affect the search result. Basically, CompRet can find a larger number of routes with a larger dataset and deeper depth setting. However, for practical applications, such a dataset and parameters should be determined adaptively based on the given case. This must be explored in future work. Further, template-based methods possess weaknesses such as computationally expensive subgraph isomorphism calculation. In the future, it may be effective to construct chemical reaction networks by using template-free methods.

This study only considered simple methods for the recommendation. Designing synthetic route evaluation metrics that are effective in all situations is a challenging task, because critical aspects of route design depend on the chemist’s objectives and/or needs. However, various route evaluation methods, including SCScore and other deep learning-based methods, have been proposed in recent years. If these methods are appropriately combined...
with CompRet, the customized framework could function as a user-friendly route recommendation system. The results obtained in this work are considered to be a successful example of bridging CASP and the field of discrete mathematics and developing an enumeration algorithm from a new perspective. This can enable the practical improvement of CASP through algorithmic techniques.

Acknowledgement

We would like to thank Dr. Nobuo Cho for the valuable discussions and comments from the perspectives of an expert synthetic organic chemist. This work was supported by MEXT as “Program for Promoting Researches on the Supercomputer Fugaku” (Application of Molecular Dynamics Simulation to Precision Medicine Using Big Data Integration System for Drug Discovery). K. Tsuda is supported by NEDO P15009, SIP (Technologies for Smart Bio-industry and Agriculture), JST CREST JPMJCR1502 and JST ERATO JPMJER1903.

Supporting Information Available

This material is available free of charge via the Internet at http://pubs.acs.org/.

References

(1) Corey, E. J. General methods for the construction of complex molecules. Pure Appl. Chem. 1967, 14, 19–38.

(2) Corey, E.; Wipke, W. T. Computer-assisted design of complex organic syntheses. Science 1969, 166, 178–192.

(3) Corey, E. J. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (Nobel Lecture). Angew. Chem., Int. Ed. Engl. 1991, 30, 455–465.
(4) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and evaluation of chemical synthesis—SECS. *Artif. Intell.* **1978**, 11, 173–193.

(5) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. *Organic Synthesis, Reactions and Mechanisms*; Springer, 1987; pp 19–73.

(6) Hendrickson, J. B. Systematic synthesis design. 6. Yield analysis and convergency. *J. Am. Chem. Soc.* **1977**, 99, 5439–5450.

(7) Wipke, W.; Ouchi, G. I.; Krishnan, S. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artif. Intell.* **1978**, 11, 173–193.

(8) Funatsu, K.; Sasaki, S.-I. Computer-assisted organic synthesis design and reaction prediction system, “AIPHOS”. *Tetrahedron Comput. Methodol.* **1988**, 1, 27–37.

(9) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **2017**, 3, 1103–1113.

(10) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **2017**, 3, 1237–1245.

(11) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **2020**, 11, 3355–3364.

(12) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **2018**, 58, 252–261.

(13) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **2017**, 3, 434–443.
(14) Schreck, J. S.; Coley, C. W.; Bishop, K. J. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* 2019,

(15) Watson, I. A.; Wang, J.; Nicolaou, C. A. A retrosynthetic analysis algorithm implementation. *J. Cheminf.* 2019, 11, 1.

(16) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model.* 2019, 59, 673–688.

(17) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. Prediction and Interpretable Visualization of Retrosynthetic Reactions using Graph Convolutional Networks. *J. Chem. Inf. Model.* 2019,

(18) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 2018, 555, 604.

(19) Coley, C. W.; Thomas, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 2019, 365, eaax1566.

(20) Satyanarayana, S. D.; Hill, R. A. Medicinal chemistry for 2020. *Future Medicinal Chemistry* 2011, 3, 1765–1786.

(21) Grzybowski, B. A.; Bishop, K. J.; Kowalczyk, B.; Wilmer, C. E. The ‘wired’ universe of organic chemistry. *Nat. Chem.* 2009, 1, 31.

(22) Kowalczyk, B.; Bishop, K. J.; Smoukov, S. K.; Grzybowski, B. A. Synthetic popularity reflects chemical reactivity. *J. Phys. Org. Chem.* 2009, 22, 897–902.

(23) Fialkowski, M.; Bishop, K. J.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and evolution of organic chemistry. *Angew. Chem., Int. Ed. Engl.* 2005, 44, 7263–7269.
(24) Bishop, K. J.; Klajn, R.; Grzybowski, B. A. The core and most useful molecules in organic chemistry. *Angew. Chem., Int. Ed. Engl.* **2006**, *45*, 5348–5354.

(25) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed. Engl.* **2016**, *55*, 5904–5937.

(26) Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem., Int. Ed. Engl.* **2012**, *51*, 7928–7932.

(27) Wender, P. A. Toward the ideal synthesis and molecular function through synthesis-informed design. *Nat. Prod. Rep.* **2014**, *31*, 433–440.

(28) Wender, P. A.; Quiroz, R. V.; Stevens, M. C. Function through synthesis-informed design. *Acc. Chem. Res.* **2015**, *48*, 752–760.

(29) Nagai, A. Application of df-pn+ to Othello endgames. Proceedings of Game Programming Workshop’99. 1999.

(30) Nagai, A. Df-pn algorithm for searching AND/OR trees and its applications. *PhD thesis, Department of Information Science, University of Tokyo* 2002.

(31) Heifets, A.; Jurisica, I. Construction of new medicines via game proof search. Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012.

(32) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning. Advances in Neural Information Processing Systems. 2019; pp 7224–7234.

(33) Chaouiya, C. Petri net modelling of biological networks. *Briefings Bioinf.* **2007**, *8*, 210–219.
(34) ChemAxon, Reactor. https://chemaxon.com/, Accessed on 2020-05-01.

(35) Riniker, S.; Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminf.* 2013, 5, 26.

(36) Landrum, G., et al. RDKit: Open-source cheminformatics. [http://www.rdkit.org/](http://www.rdkit.org/), Accessed on 2020-05-01.

(37) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* 2015, 7, 20.

(38) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* 2015, 55, 39–53.

(39) Maaten, L. v. d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 2008, 9, 2579–2605.

(40) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., et al. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 2011, 12, 2825–2830.

(41) Reaxys. [https://www.reaxys.com/](https://www.reaxys.com/), Accessed on 2020-05-01.

(42) Enamine. [https://enamine.net/building-blocks](https://enamine.net/building-blocks), Accessed on 2020-05-01.

(43) Savage, C. A survey of combinatorial Gray codes. *SIAM review* 1997, 39, 605–629.

(44) Marino, A. *Analysis and enumeration: algorithms for biological graphs*; Springer, 2015; Vol. 6.

(45) Gansner, E.; Koutsofios, E.; North, S. Drawing graphs with dot. [https://graphviz.github.io/_pages/doc/info/lang.html](https://graphviz.github.io/_pages/doc/info/lang.html), 2006; Accessed on 2020-05-01.
(46) WebMD. https://www.webmd.com/drugs/2/drug-12065/cetirizine-oral/details, Accessed on 2020-05-01.

(47) Saunders, J. Top drugs: top synthetic routes; Oxford University Press, 2011.

(48) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. Acc. Chem. Res. 2018, 51, 1281–1289.

(49) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to predict chemical reactions. J. Chem. Inf. Model. 2011, 51, 2209–2222.

(50) Kayala, M. A.; Baldi, P. ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. J. Chem. Inf. Model. 2012, 52, 2526–2540.

(51) Soh, S.; Wei, Y.; Kowalczyk, B.; Gothard, C. M.; Baytekin, B.; Gothard, N.; Grzybowskia, B. A. Estimating chemical reactivity and cross-influence from collective chemical knowledge. Chem. Sci. 2012, 3, 1497–1502.

(52) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. J. Chem. Inf. Model. 2020,
CompRet, a recommendation framework for synthesis planning

Target

Recommendation

CompRet

Enumeration Visualization