Substructure-based Neural Machine Translation for Retrosynthetic Prediction †

Umit V. Ucak, ‡ Taek Kang, ¶ Junsu Ko, § and Juyong Lee*, ‡

‡Division of Chemistry and Biochemistry, Department of Chemistry, Kangwon National University, Chuncheon, 24341, Republic of Korea
¶Center for Neuro-Medicine, Korea Institute of Science and Technology (KIST), Seoul, 02792, Republic of Korea
§Arontier co., Seoul, 06735, Republic of Korea

E-mail: juyong.lee@kangwon.ac.kr

Abstract

With the rapid improvement of machine translation approaches, neural machine translation has started to play an important role in retrosynthesis planning, which finds reasonable synthetic pathways for a target molecule. Previous studies showed that utilizing the sequence-to-sequence frameworks of neural machine translation is a promising approach to tackle the retrosynthetic planning problem. In this work, we recast the retrosynthetic planning problem as a language translation problem using a template-free sequence-to-sequence model. The model is trained in an end-to-end and a fully data-driven fashion. Unlike previous models translating the SMILES strings of reactants and products, we introduced a new way of representing a chemical reaction based on molecular fragments. It is demonstrated that the new approach yields better prediction results than current state-of-the-art computational methods. The

†A footnote for the title
new approach resolves the major drawbacks of existing retrosynthetic methods such as generating invalid SMILES strings. Specifically, our approach predicts highly similar reactant molecules with an accuracy of 57.7%. In addition, our method yields more robust predictions than existing methods.

1 Introduction

Although knowledge in organic chemistry has accumulated over decades, designing an efficient synthetic route for a target molecule remains a crucial task in organic synthesis. The retrosynthetic approach suggests a logical synthetic route to generate a target molecule from a set of available reactants and reagents. This approach is both iterative and recursive in nature since a sequential computation of retrosynthetic transformation is required. Retrosynthetic transformation occurs recursively until much simpler and commercially available molecules are identified.

Computational retrosynthetic analysis initially formalized in 1969 by Corey and Wipke in an algorithmic manner. The algorithm considers all possible disconnections with known reaction types, which reduce the complexity of a product and progress until chemically reasonable pathways are identified. Such disconnections were based on handcrafted minimal transformation rules known as reaction templates. Manual encoding of those transformation rules necessitates deep chemical expertise and intuition. Manual management of synthetic knowledge is a highly complicated task considering a large number of transformation rules (>10000) that must be hand-coded. Furthermore, being dependent on reaction templates potentially limits prediction accuracy, particularly if a reaction is outside of the template domain. Later studies offer valuable help to chemists in finding better routes faster by enabling automated extraction of reaction templates. However, they do not address the above-mentioned limitations inherited from their precedents. Computer-aided synthesis planning has been well summarized in many recent reviews.

Reaction Predictor developed by Kayala et al. was the first template-free approach. It
was a mechanistic level of strategy that merges the idea of rule-based modeling and machine learning within its framework. Jin et al.\textsuperscript{27} proposed a novel template-free, entirely data-driven approach based on the Weisfeiler-Lehman networks.\textsuperscript{28} Both approaches provide end-to-end solutions to generate candidate products. Theoretical findings provided by Cadeddu et al.\textsuperscript{29} have further motivated the development of other template-free methods for the forward- or retro-reaction prediction tasks using various types of neural machine translation (NMT) architectures.\textsuperscript{30–38} Based on an explicit analogy between sentences in a language corpus and molecules in a chemical corpus, i.e. chemical space, Cadeddu et al. showed that the rank-frequency distributions of substructures as the building blocks of molecules are similar to those of words in a natural language corpus. This verification implies that the concepts of linguistic analysis are readily applicable to tackle the problems of forward- and retro-reaction prediction. In this context, a retrosynthetic prediction is appropriate for applying the sequence-to-sequence framework\textsuperscript{39–41} of machine translation.

Sequence-to-sequence learning uses a recurrent neural network (RNN) layer to map a source sequence of an arbitrary length into a fixed dimensional context vector consisting of real numbers. The context vector contains information about the syntactic and semantic structure of the source sequence. In connection with this RNN layer, another RNN decodes the context vector to a target sequence. In this regard, the two RNN units together act like a pair of encoder-decoder system. Sutskever et al.\textsuperscript{41} showed that long short-term memory (LSTM)\textsuperscript{42}-based architectures can solve general sequence-to-sequence problems because of their ability to handle long-range relations in sequences. Liu et al.\textsuperscript{34} proposed the first multi-layered LSTM-based sequence-to-sequence model for retrosynthetic prediction. Its gated recurrent unit (GRU)\textsuperscript{39} variant was proposed by Nam and Kim\textsuperscript{32} for the forward reaction prediction.

Recently, the best performing NMT models include an attention mechanism\textsuperscript{40,43} as a part of their neural architectures to enhance their performances on longer sentences.\textsuperscript{27,32–34} There are also retrosynthetic predictors built on the Transformer architecture,\textsuperscript{31,37,44,45} based solely
on the attention mechanism. Encoder-decoder models, especially once an attention mechanism is introduced, all employ similar strategies to handle a translation task. The SMILES representations of molecular structures are typical inputs for the sequence-to-sequence based models. However, none of the previously reported models has focused on translation at a substructural, fragment, level.

In this paper, we propose a template-free approach for retrosynthetic reaction prediction by learning the chemical change at a substructural level. Our approach represents a molecule as a sentence based on a set of substructures corresponding to a word by using the MACCS keys. We also present a unique tokenization scheme that properly eliminates problematic issues originate from SMILES-based tokenization. Our model consists of bidirectional LSTM cells, and is trained in a fully data-driven and end-to-end fashion without prior reaction class information. We thoroughly discuss all the aspects of our methodology, including dataset and descriptor curation steps. Evaluation results are presented based on three datasets derived from the United States Patent and Trademark Office (USPTO) reaction dataset.

This paper is organized as follows. In Section 2, we suggest a new way of tokenization followed by curation together with the analysis of the dataset and descriptor. We briefly describe the model architecture and evaluation procedure for accuracy calculations. In Section 3, the results of a set of translation experiments are discussed with an emphasis on the benefits of the MACCS key-based molecular representation. Finally, the strengths and limitations of our approach are discussed in Section 4.

2 Method

2.1 Dataset

In this study, we used the filtered US patent reaction dataset, USPTO, which is obtained with a text-mining approach. Schwaller et al. eliminated the duplicated reaction strings in

4
the dataset without atom-mapping. They also removed 780 reactions due to SMILES canonicalization failures with RDKit.\textsuperscript{50} The inherent limitation of the data is that the vast majority of entries are single product reactions. Thus, only single product cases corresponding to 92\% of the dataset are used in this study.

The SMILES line notation\textsuperscript{51} represents molecular structures as a linear sequence of letters, numbers, and symbols. Hence, from a linguistic perspective, SMILES can be regarded as a language with grammatical specifications. However, in our approach, molecules are represented as a set of fragments using the MACCS keys consisting of 166 pre-defined substructures.\textsuperscript{46} This binary bit-based molecular descriptor converts a molecule into a 166 bit vector, in which each bit indicates the presence of a feature taken from a predefined dictionary of SMARTS patterns.\textsuperscript{52}

2.2 Descriptor Curation

In our approach, a molecule is represented as a set of fragments using the MACCS keys. The number of occurrences of each MACCS key in our dataset was investigated. Also, we compared the results obtained for one million randomly sampled drug-like small molecules, a subset of the Generated Data Base-13 (GDB-13) consisting of 975 million molecules.\textsuperscript{53,54} Figure 1 shows the normalized frequency distributions of the MACCS keys on both databases. A direct pairwise comparison rationalizes reducing the number of MACCS keys (Figure 1). In this study, five keys that never occurred and nine keys that are not frequently observed in the USPTO database are omitted. Based on the comparison, additional 26 keys that are never or hardly ever observed in the GDB-13 database are also excluded.

It is not essential to use all MACCS keys because our main interest is to develop a model to analyze drug-like molecules. Removing redundant keys based on the occurrence analysis has apparent advantages. It shortens the lengths of source and target sentences and provides a better rank distribution of the keys used in the translation process. In our approach, every molecule is represented by 126 MACCS keys, which are able to represent 98\% of the
Figure 1: Descriptor curation based on the rate of occurrences. Filtered US patent reaction dataset and 1 million randomly sampled drug-like small molecules as a subset of the enumerated database (GDB-13) are compared to investigate the MACCS keys probability distribution profiles.

one million randomly sampled subset of GDB-13 adequately. In machine translation tasks that chemists are dealing with, source and target molecules are placeholders corresponding to reactants and products interchangeably. The selection is dependent on the intended analysis. For a retrosynthetic prediction task, source and target sentences refer to products and reactants, respectively.

2.3 Reaction Preprocessing

Our model considers only non-zero indices of curated MACCS keys. English letters were assigned to the ranked non-zero MACCS keys based on their rank-frequency distribution to form unique artificial "words". Our product and reactant sentences consist only of 126
words, i.e., each sentence is a sorted list of independent fragments. Single-lettered words were generated by using the upper- and lower-cases of the most frequent 21 letters in English. Double-lettered words were constructed by concatenation of “x” and “z” for every 42 single letter, which allowed us to cover all 126 MACCS keys. Namely, our lettered fragment vocabulary has a fixed length of 126. The generation process of an example product-reactant pair is illustrated in Table 1. The same procedure was applied to all reactions of the dataset. The complete mapping of the MACCS keys to artificial words is listed in Supplementary Information.

Table 1: Data preparation procedure to obtain product and reactant sentences for a retrosynthetic prediction task.

| Step                                         | Example (entry # 20032, Lowe’s USPTO dataset$^{48}$) |
|----------------------------------------------|------------------------------------------------------|
| **Reaction**                                 |                                                      |
| Reaction SMILES (Canonicalized)              | Reactants > Reagents · Solvents > Products           |
| (Canonicalized)                              | C=CC.CC(=O)O>CC(=O)OC(C)=O.O=O>CC(=O)OCC(C)OC(C)=O |
| Removal of reagents and solvents             | [‘C=CC’, ‘CC(=O)O’], ‘CC(=O)OCC(C)OC(C)=O’          |
| MACCS key domain *                          | [34, 99, 160], [123, 139, 154, 157, 159, 160, 164], |
| (Nonzero keys)                               | [72, 108, 109, 115, 116, 123, 126, 132, 136, 140, 141, 146, 149, 152, 153, 154, 155, 157, 159, 160, 164] |
| Letter assignment (Frequency based)         | [‘Uz’, ‘dz’, ‘s’], [‘Y’, ‘Ex’, ‘c’, ‘r’, ‘h’, ‘s’, ‘t’], |
|                                              | [‘iz’, ‘Fx’, ‘fx’, ‘ez’, ‘Gx’, ‘Y’, ‘E’, ‘hx’, ‘ux’, ‘Hx’, ‘Ix’, ‘O’, ‘H’, ‘v’, ‘u’, ‘c’, ‘w’, ‘r’, ‘h’, ‘s’, ‘t’] |
| Product                                      | iz Fx fx ez Gx Y E hx ux Hx Ix O H v u c w r h s t |
| Reactant(s)                                  | Y Ex c r h s t - Uz dz s                            |

$^{*}$ Original MACCS key indices (RDKit implementation)
2.4 Reaction Dataset Curation

The product-reactant pair dataset was further curated before being processed by our translation machine. After representing every molecule with the 126 truncated MACCS keys, a series of filters were applied to remove identical product-reactant pairs and internal twins. Internal twins are the pair of data entries whose product and reactant sentences are identical. They appeared whenever the chemical changes were beyond the sensitivity of our MACCS key-based representation. Because we associate molecules with MACCS keys to operate on a substructural subspace, a certain amount of information is lost. Our preprocessing procedure resulted in 5748 internal twins, and they are removed from our dataset. In addition, the reactions with three or more reactants were excluded. The length of the longest pair was set to 100 to avoid lengthy fragment sequences, as shown in Figure 4.

The product-reactant pairs were then put into an injective map generator to guarantee one-to-one correspondence between product and reactant sentences. If a reactant sentence is composed of two reactants, we sorted them in descending order according to their sequence length. Reactants were separated by the "–" sign. The curated dataset, containing a total of 352,546 product-reactant pairs, was further subdivided by the number of the reactant molecules in each pair into two disjoint subsets: single reactant and double reactant datasets. Organizing the dataset in this manner was essential to assess model performance independently. The dataset curation steps along with the dataset sizes are summarized in Scheme 1.

2.5 Model Architecture

Our sequence-to-sequence neural network comprises two bidirectional LSTMs: one for an encoder and the other for a decoder. Besides, we used unidirectional LSTMs to quantify the improvement in model’s performance with the use of bidirectional LSTMs. The encoder and decoder layers were connected through Luong’s global attention mechanism, which captures non-local relations between all elements of source sequences. The attention mechanism allows
Scheme 1: Dataset curation process and obtaining training/test pairs. P: Product, R: Reactant. Details to the different steps are given in the text.
neural networks to focus on different parts of a source sentence, and to consider non-linear relationships between words during a training process. The global attention mechanism used in this study, in essence, is similar to the first attention mechanism suggested by Bahdanau et al.,\textsuperscript{40} for machine translation tasks. The global approach focuses the ”attention” on all the words on the source sentence to compute a global context vector for each target word at each time step in the decoder unit. Therefore, the global context vector represents the weighted sum over all the source hidden states. This context information leads to improved prediction accuracy.

2.6 Training Details

Our curated datasets were randomly split into 9:1 to generate training and testing sets. The validation sets were randomly sampled from training sets (10%). The word embeddings were used to represent lettered fragments in the vocabulary. After the embedding layer was created, a trainable tensor holding 126-dimensional fixed-length dense vectors was randomly initialized. A method of embedding class then accessed the embedding of each word through a lookup on the tensor. We used the stochastic gradient descent algorithm\textsuperscript{56} to train all parameters of the encoder-decoder model. The cross-entropy function was used as a loss function.

For each dataset, we performed a series of tests within the range of hyper-parameter space as described in the Supporting Information (Table 6) to achieve optimal performance. Based on the preliminary experiments, we generated an encoder and a decoder with two Bi-LSTM layers containing 2000 hidden units at each layer. A dropout layer with a dropout rate of 0.1 was included following the hidden layer to avoid overfitting. To avoid a potential exploding gradient problem, we introduced gradient clipping\textsuperscript{57} to guarantee that the norm of the gradients did not exceed a threshold (0.25) during backpropagation. The initial learning rate was set to 4.0, and it decayed with a factor of 0.85 every three epochs.\textsuperscript{33}

With these hyper-parameters, the average training speed was approximately 3300 words
per second, with a batch size of 64 on a single NVIDIA RTX 2080Ti GPU card. Larger batch sizes were not tested due to memory constraints, which likewise apply to the hidden layer’s size. We trained our models for a minimum of 30 epochs, and each epoch took about 2 hours for the curated dataset consisting of 320K sentence pairs. The details of our key hyper-parameters are available in the Supporting Information, Table 6.

Our model was implemented in Python version 3.6.8 together with PyTorch\textsuperscript{58} version 1.3.0. The open-source RDKit module version 2020.03.1\textsuperscript{50} was utilized to obtain MACCS keys and similarity maps.\textsuperscript{59}

### 2.7 Evaluation Procedure

To evaluate the performance of our retrosynthetic model, the Tanimoto coefficient was selected as a similarity metric, which is identified as one of the best metrics to compute structural similarity.\textsuperscript{60} Pairwise similarities between the predicted sequences and ground truth of all test molecules were calculated. Tanimoto coefficient ($T_c$) measured between two chemical structures have a value between 0 and 1. The coefficient is zero if molecules share no common fragments while identical molecules have a Tanimoto coefficient of unity. Though these are the cases for the two ends of the Tanimoto similarity metric, there is no single criterion that defines similar and non-similar molecules. We defined three threshold values (0.50, 0.70, and 0.85) to assess the quality of translation experiments. The similarity between predicted and ground truth sentences was computed at the end of each epoch for every pair appear in the validation set using the Tanimoto similarity measure (Equation 1).

$$T_c(R, P) = \frac{\sum_i R_i P_i}{\sum_i (R_i)^2 + \sum_i (P_i)^2 - \sum_i R_i P_i}$$ \hspace{1cm} (1)

Our machine yields predictions either with one or two reactants as all reactions are
contained in the combined dataset. There are thus multiple possibilities for comparing predicted sequences with ground truths. The potential pairs for evaluation corresponding to the number of reactants are listed in Table 2. Tanimoto similarities between all possible pairs of predicted sequences and ground truths were calculated. Then, the pair(s) with the highest similarity was selected based on the assumption that more similar structures are more likely to be matched.

Table 2: The possible pairs between predicted sequences and ground truths are presented. The similarity of each pair is computed with the Equation 1.

| Ground Truth | Predictions | List of possible pairs |
|--------------|-------------|------------------------|
| $P \rightarrow R_A + R_B$ | $P \rightarrow P_A + P_B$ | $(R_A P_A - R_B P_B), (R_A P_B - R_B P_A)$ |
| $P \rightarrow P_C$ | $P \rightarrow P_C$ | $R_A P_C, R_B P_C$ |
| $P \rightarrow R_C$ | $P \rightarrow P_A + P_B$ | $R_C P_A, R_C P_B$ |
| $P \rightarrow P_C$ | $P \rightarrow P_C$ | $R_C P_C$ |

3 Results and Discussion

3.1 Prediction Accuracy

The performance of our model was assessed based on three datasets: single reactant, double reactant, and the combined test set. Evaluation results on the test sets are summarized in Table 3. The quality of predictions of each test dataset is expressed in terms of pairwise Tanimoto similarity values. We introduced three criteria for evaluating the success rates of our translation models: 1) the number of exact matches ($T_c = 1.0$), 2) the number of bioactively similar matches ($0.85 < T_c < 1.00$) and 3) the overall success rate presented as the average Tanimoto similarity between predicted and true sequences (a series of fragments) over all the test molecules.

For the single reactant reactions, our bidirectional-LSTM model achieved an accuracy
Table 3: Success rate over single reactant, double reactant and combined test sets based on the Tanimoto similarity metric.

|                  | Single | Double | All    |
|------------------|--------|--------|--------|
| **Size**         |        |        |        |
| Training pairs   | 88,151 | 229,141| 317,292|
| Test pairs       | 9,794  | 25,460 | 35,254 |
| Test molecules   | 9,794  | 50,911 | 55,958 |
| Average pair length | 74    | 73     | 74     |
| **Success rate** |        |        |        |
| Bi-LSTM          |        |        |        |
| $T_c = 1.0$      | 29.0%  | 27.9%  | 25.3%  |
| $0.85 < T_c < 1.00^a$ | 28.7%  | 10.5%  | 12.9%  |
| $\bar{T_c}^b$   | 0.84   | 0.66   | 0.68   |
| LSTM             |        |        |        |
| $T_c = 1.0$      | 22.9%  | 21.6%  | 19.4%  |
| $0.85 < T_c < 1.00^a$ | 29.7%  | 10.2%  | 12.5%  |
| $\bar{T_c}^b$   | 0.82   | 0.62   | 0.64   |

$^a$ Bioactively similar molecules. $^b$ Average similarity.

of 57.7% based upon the combined use of the first two criteria. The percentages of exact and bioactively similar matches were 29.0% and 28.7%, respectively. The average $T_c$ value between predicted and true sequences was 0.84. These results demonstrate that our machine predicts single reactant reactions with high accuracy. For the double reactant reactions, the success rate of the exact matches (27.9%) was almost identical to that of the single reactant reactions. However, the success rate of highly similar predictions deteriorated to 10.5% from 28.5%. For the combined set, 25.3% of predictions were accurate, and 12.9% of them were highly similar. Similarly, the average $T_c$ values dropped from 0.84 to 0.66 and 0.68 for datasets containing double and combined reactants.

One reason for the worse accuracy of the double and combined sets is that the "-" sign should be appropriately predicted. Another reason is the frequent occurrence of small
molecules represented with a small number of MACCS keys in these datasets. In fact, 477 molecules represented with less than 7 MACCS keys appeared in 61822 different reactions. To be more specific, 3944 reactions contain a reactant represented with one of the seven MACCS keys described in Figure S6. The number of unique structures corresponding to those keys was, however, only 29. Because such small and simple structures were dense in these datasets, wrongly predicted fragments contributed significantly (a value of zero in 1-bit cases) to the success rate.

Our result also demonstrates that the bidirectional LSTM-based model outperforms the unidirectional LSTM-based model. The success rates of exact matches become lower by about 6% for all the datasets consistently. This is possibly due to the fact that our MACCS key-based representation of a molecule does not depend on the order of keys. In other words, most information about molecules and chemical reactions are embedded into the co-occurrences of keys.

We compared the prediction accuracy of our approach with other retrosynthetic prediction methods without considering reaction class labels because no prior reaction class information was provided to our model. Several recent reports summarized the prediction accuracy of various models.\textsuperscript{37,61} According to reproduced results presented by Lin et al.,\textsuperscript{37} top-1 accuracy ranges from 28.3\% (Liu et al.\textsuperscript{34} LSTM model over the USPTO 50K dataset) to 54.1\% (Transformer model over the USPTO MIT dataset by Lin et al.\textsuperscript{37}). As an alternative approach, Coley’s similarity-based model\textsuperscript{62} achieved a top-1 accuracy of 37.3\% on the USPTO 50K dataset.

To identify how our model learns the grammar of chemical reactions, the evolution of prediction accuracy with respect to threshold values along training epoch for the single reactant validation set is illustrated in Figure 2. In particular, it is demonstrated that the network successfully learned reaction rules by capturing the alterations of molecules at a substructural level. The number of exact matches ($T_c = 1.0$) increased rapidly during the first 10 epochs. After 20 epochs, the value became almost tripled. The likelihood of
making a better prediction for each fragment becomes higher during training. This is a clear
indication of successful training. The improvement in exact matches appears to be a result
of the respective declines in non-exact matches except extremely bad predictions ($T_c < 0.50$).
The quality of bad predictions (ca. 5% of the validation set) did not improve probably due
to the insufficient information, complexity, and noise contained in the data. This observation
similarly repeated for all the other datasets.

![Evaluation of Single Reactant Retrosynthetic Prediction Model](image)

Figure 2: Number of matches at different ranges of similarity. $T_c$ refers to Tanimoto similarity
coefficient.

In this study, we assumed that candidate reactants with $T_c > 0.85$ are similar enough to
their true counterparts. To validate this assumption, we assessed the quality of candidate
reactants by comparing them with true reactants. We investigated whether the following
factors were correct: functional group interconversion (FGI) or bond disconnection, reactive
functional groups, and core structures. The accuracy of side-substituents is regarded as less
significant for matching the reactants’ functionality, especially when they are simple alkyls.
Randomly chosen predictions exemplifying possible prediction cases are presented in Table 4. Similarity maps are presented to visualize similarities between candidates and true reactants.

Table 4: Non-exact candidates varied in their degree of similarity are presented. Similarity score calculations and similarity maps using the Morgan fingerprints and the Tanimoto metric are shown. Colors indicate atom-level contributions to the overall similarity (green: increases similarity score, red: decreases similarity score, uncolored: has no effect)

| Product | True Reactant(s) | Candidate Reactant(s) | Similarity Map(s) | $T_c$ |
|---------|------------------|-----------------------|-------------------|------|
|         |                  |                       |                   |      |
| **Single reactant reaction : non-exact candidate** |         |                       |                   |      |
| 1       | ![Molecule Image](image1) | ![Molecule Image](image2) | ![Similarity Map Image](image3) | 0.64 |
| 2       | ![Molecule Image](image4) | ![Molecule Image](image5) | ![Similarity Map Image](image6) | 0.85 |
| **Double reactant reactions : two non-exact candidates** |         |                       |                   |      |
| 3       | ![Molecule Image](image7) | ![Molecule Image](image8) | ![Similarity Map Image](image9) | (0.84, 0.30) |
| 4       | ![Molecule Image](image10) | ![Molecule Image](image11) | ![Similarity Map Image](image12) | (0.94, 0.81) |
| **Double reactant reactions : exact and non-exact candidates** |         |                       |                   |      |
| 5       | ![Molecule Image](image13) | ![Molecule Image](image14) | ![Similarity Map Image](image15) | (0.68, 1.00) |
| 6       | ![Molecule Image](image16) | ![Molecule Image](image17) | ![Similarity Map Image](image18) | (1.00, 0.46) |
Reaction 1 resulted in a reactant where the main chain composed of 8 carbons, and an α,β-unsaturated aldehyde group in the correct position was derived accurately (Table 4). Although an ester was expected rather than an aldehyde, an aldehyde reduction could also provide the same target alcohol. This indicates that our prediction identified the functional group interconversion correctly. On the other hand, one olefin was missing and the position and number of two methyl groups out of four were misinterpreted. In reaction 2, aside from the location of an ester group, core heterocyclic rings, pyridine and thiazole, and their connections were accurately generated. In the true reactant, a methyl ester group was attached to C6 of pyridine, whereas the ethyl ester group was attached to C4 of the thiazole ring in our candidate. If the position of the ethyl ester group was accurate, it would require a single-step reduction to obtain alcohol group. In reaction 3, the core structure of pyrazole ring and its methyl ester group were predicted accurately. However, there was no chloride, one of the reactive functional groups, and substituents on the pyrazole ring as well as structure of thiol were misinterpreted.

The result of reaction 4 showed that our model correctly predicted the core structures, bond disconnections, and reactive functional groups. However, the number and position of halides were wrong. In the case of reaction 5, one reactant was predicted precisely, but the other was partially incorrect. In wrongly predicted candidate, a (phenyl)methyl group appeared instead of a (2-naphthyl)vinyl group, but the reactive functional group, acylhydrazine, was correctly produced. The result of reaction 6 revealed the exact match for N-Hydroxyphthalimide as a precursor for O-hydroxylamine. However, the structure of the alkyl halide lacked a phenylene group. The core structure estimation failed to a great extent for this reaction. On the other hand, the reactive functional groups and bond disconnection are suggested correctly.

The quantitative summary of the assessment above is given in Table 5. The three criteria: functional group interconversion or bond disconnection, core structure, and reactive functional group are weighted equally. They are utilized to form a chemically reasonable
score along with similarity scores. The evaluation was carried out by following procedure. First, we identified less significant parts of candidate molecules by comparing them with the product and true reactants. Second, core structures were identified; true reactants were separated into fragments, e.g., functional group, chain, ring. Afterwards, each fragment of a candidate molecule was evaluated against fragments found in second step in terms of the core structure, type and positions of side-substituents in an equally weighted manner. Finally, equal weight was given to the correctness of fragments’ positions within candidate reactants. Concerning the core structure, the longest chain of carbons and/or a ring, either of which may possess heteroatoms such as O, N, S, were taken into account together with important side-substituents and their positions. Because functional group interconversion or bond disconnection as well as reactive functional groups are the most significant factors of retrosynthetic analysis, the correct positions of reacting sites are scored strictly as true/false values corresponding to 1 and 0, respectively. We scored each candidate reactant individually and averaged the results to obtain a final score for each criterion.

It is noticeable that our model correctly predicted functional group interconversion or bond disconnection of all six reactions. Except for reaction 3, reactive functional groups are correctly reflected. We observe that prediction errors that affect the score are mainly associated with core structures. We applied this knowledge-based scoring strategy to a more specific set containing ten randomly chosen reactions where candidate reactants, on average, lies within bioactively similar region ($T_c = 0.87$) (Figure S5 and Table S7). The results clearly show that our model is highly accurate in predicting functional group interconversion or bond disconnection as well as reactive functional group for bioactively similar reactant candidates. A similar argument can also be made regarding the prediction errors, since they mainly originate from core structures.

The chemical inspection of reactions indicates that average similarity scores and knowledge-based scores are closely related. Our scoring approach offers a clear idea about the quality of candidate reactants and similarity scores are in good agreement with those manually in-
Table 5: Summary of quality assessment of candidate reactants. The functional group interconversion (FGI) or bond disconnection and reactive functional group columns represent the correctness in a True(1)/False(0) fashion. The core structure column presents the averaged accuracy of the core structures of candidate molecules by capturing the correctness of core structures themselves as well as the type and positions of side-substituents. The source of errors are given inside the parenthesis e.g., "C2=0.33, 2/3 fragments" implies that the accuracy of candidate reactant 2 is 0.33 because 2 out of 3 fragments are wrongly predicted. The average of the three criteria is given as a separate column. Tc column represents the averaged Tc values of candidate reactants. C1: Candidate 1, C2 : Candidate 2.

| Reaction Number | FGI or Bond Disconnection | Core Structure          | Reactive Functional Group | Avg. Tc |
|-----------------|---------------------------|-------------------------|---------------------------|---------|
| 1               | 1.00                      | 0.33 (2/3 fragments)    | 1.00                      | 0.78    | 0.64 |
| 2               | 1.00                      | 0.67 (1/3 fragment’s positions) | 1.00  | 0.89 | 0.85 |
| 3               | 1.00                      | 0.69 (C1=0.88, fragment’s side subst.; C2=0.5, 1/2 fragments) | 0.50 (1 for thiol, 0 for chloride) | 0.73    | 0.57 |
| 4               | 1.00                      | 0.96 (C1=0.92, position of side subst.; C2=1.0, Cl is omitted) | 1.00  | 0.99 | 0.86 |
| 5               | 1.00                      | 0.83 (C1=0.67,1/3 fragments; C2 is exact) | 1.00 | 0.94 | 0.84 |
| 6               | 1.00                      | 0.67 (C1 is exact; C2=0.33, 2/3 fragments) | 1.00 | 0.89 | 0.73 |

Similarity measurements yield lower scores than knowledge-based scores possibly due to the inclusion of side chains and geometrical factors (more detailed topological exploration is provided by Morgan fingerprint). Although the interpretation of the similarity score is rather difficult to assess objectively, it can be used for assessing the quality of retrosynthetic predictions. Higher similarity scores indicate that the desired molecules are more synthetically accessible according to the rules of organic chemistry.

### 3.2 Advantages of Our Model

The key advantage of our word-based MACCS keys model over the character-based SMILES methods is that the network needs to learn relatively simpler grammatical rules: ascending
order and co-occurrence of keys, to yield meaningful results. In the SMILES-based methods, a network has to comprehend not only the complicated grammar of SMILES but also the canonical representation to predict synthetically correct sequences. As summarized by Liu et al., the difficulty of learning the syntactic structure of SMILES notation possibly causes problematic outcomes such as invalid SMILES strings. In general, existing character-based models suffer from the generation of literally invalid, literally valid but chemically unreasonable, or literally and chemically valid but unfeasible candidates. We avoided this problem by projecting the SMILES representation of a molecular structure into a substructural domain. Our approach can be an effective solution to these technical problems at a fundamental level.

In general, the likelihood of making correct retrosynthetic predictions remains rather low. Indeed, the accuracy of retrosynthetic planning tasks is twice as much lower than the level of accuracy achieved at forward reaction prediction tasks. This is especially true assuming that several possible synthetic routes are available for the forward reaction. It is worth noting that the content of the dataset used in the reverse mapping, could also be responsible for the network’s behavior. Mapping a reactant from a reactant domain to a product domain and then reversing it does not necessarily produce the original reactant considering the level of abstraction used to describe the molecules in our dataset. There is a chance that the presence of one-to-many mappings from a product to a reactant domain may create confusion during the learning process. Equipped with these observations, a simple idea is adopted to assure a stronger pairwise functional relationship between the domains. To achieve this, we identified all one-to-many mappings and collapsed them into an injective mapping (see Scheme 1, Section 2.4) by selecting the molecule with the shortest sequence length (presumably the reactants with the lowest level of structural complexity).

Notably, our model yields robust predictions. For each independent run of the same input molecule, our model gives the same output consistently. This robustness of our model may be due to the low complexity and good interpretability of our molecular descriptor. Generally, retrosynthetic models have employed the top-N accuracy score to assess overall model per-
formances.\textsuperscript{11,34–37,45,62} However, as recently discussed by Schwaller,\textsuperscript{38} top-N accuracy score may not be an adequate metric for assessing retrosynthetic models because with each suggestion, the model tends to yield expected answers from the dataset rather than making chemically more meaningful predictions. Although MACCS keys have been criticized for their poor performance on similarity benchmarks,\textsuperscript{63} an advantage of such descriptor is that there is an one-to-one correspondence between a bit and a substructure compared to fingerprints obtained by an exhaustive generation algorithm followed by a hashing procedure. Thus, MACCS keys were a natural choice to test the proof-of-concept level of our translation methodology.

Using substructure-based representation makes it hard to convert predicted sequences into molecules. The SMILES representation of a molecular structure is not unique, but it is invertible. On the contrary, the MACCS key representation exhibits unique, non-invertible characteristics. For this reason, we need to do a lookup to access molecular structures in a pre-computed database of MACCS keys. We constructed a lookup table from the USPTO reaction dataset to associate predicted MACCS keys to reactant molecules. Also, a candidate reactant can be retrieved from a database of commercially available chemicals or a reaction database based on a similarity search. Several online chemical databases such as PubChem\textsuperscript{64} also serve the same purpose by allowing for MACCS keys-based queries.

Figure 3 depicts the seven candidates for the first reactant of the reaction 4 in Table 4 retrieved from the USPTO reaction dataset. All of the seven candidates are associated with different reactions in the database. The MACCS key representation of the retrieved molecules are identical. This implies that it is possible to find more than one match corresponding to the predicted sequence. These closely related analogs can be ordered by computing the Tanimoto coefficients using path-based or circular fingerprints as they will be different for the same set. For this purpose, we used the circular fingerprint\textsuperscript{65} with radius 2 as a bit vector. We selected the molecule with the highest similarity value among candidates as our final result.
4 Conclusion

We developed a sequence-to-sequence NMT model to extract the reaction rules of a chemical reaction automatically by learning the relationships at substructural level. By constructing an abstract language with a small size fixed-length vocabulary of non-zero elements of MACCS keys, three conceptual problems are addressed and resolved jointly: (1) erratic predictions: SMILES-based representation makes model outcomes prone to error, (2) synthetic availability: predicted molecules may not be synthetically accessible, and (3) top-N accuracy metric: suggestions made by the model may vary by model run. The comparison and quality inspections showed that our method successfully produced candidate reactants within a region $0.85 < T_c \leq 1.00$, achieving a high level of overall accuracy, particularly at functional group interconversion or bond disconnections and reactive functional groups. We believe that this proposed approach has a high potential for broad applications in organic chemistry. For the future version, it is essential to develop a better defined structural key suitable for reaction prediction purposes.
Acknowledgement

References

(1) Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; Wood, A. Organic synthesis provides opportunities to transform drug discovery. Nat. Chem. 2018, 10, 383–394.

(2) Corey, E. J. Robert Robinson lecture. Retrosynthetic thinking - Essentials and examples. Chemical Society Reviews. 1988; pp 111–133.

(3) Corey, E. J.; Cheng, X. M. The Logic of Chemical Synthesis; Wiley, 1989.

(4) Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbonic Molecules (Nobel Lecture). Angew. Chem. Int. Edit. 1991, 30, 455–465.

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

(6) Pensak, D. A.; Corey, E. J. LHASA—Logic and Heuristics Applied to Synthetic Analysis. 1977, 1–32.

(7) Salatin, T. D.; Jorgensen, W. L. Computer-assisted mechanistic evaluation of organic reactions. 1. Overview. J. Org. Chem. 1980, 45, 2043–2051.

(8) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. A collection of computer methods for synthesis design and reaction prediction. Recl. Trav. Chim. Pay-b. 1992, 111, 270–290.

(9) Fick, R.; Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Design of Syntheses for Heterocyclic Compounds. Heterocycles 1995, 40, 993–1007.

(10) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Angew. Chem. Int. Edit.; 2016; Vol. 55; pp 5904–5937.
(11) Segler, M. H.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chem. Eur. J. 2017, 23, 5966–5971.

(12) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. J. Chem. Inf. Comp. Sci. 1995, 35, 34–44.

(13) Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. J. Chem. Inf. Comp. Sci. 1999, 39, 316–325.

(14) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. J. Chem. Inf. Model. 2009, 49, 593–602.

(15) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. Org. Process Res. Dev. 2015, 19, 357–368.

(16) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. ACS Cent. Sci. 2016, 2, 725–732.

(17) 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.

(18) Segler, M. H.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. Chem. Eur. J. 2017, 23, 6118–6128.

(19) Ott, M. A.; Noordik, J. H. Computer tools for reaction retrieval and synthesis planning in organic chemistry. A brief review of their history, methods, and programs. Recl. Trav. Chim. Pay-b. 1992, 111, 239–246.
(20) Todd, M. H. Computer-aided organic synthesis. Chem. Soc. Rev. 2005, 34, 247–266.

(21) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-aided synthesis design: 40 years on. Wiley Interdiscip. Rev. Comput. Mol. Sci. 2012, 2, 79–107.

(22) Warr, W. A. A short review of chemical reaction database systems, computer-aided synthesis design, reaction prediction and synthetic feasibility. Mol. Inform. 2014, 33, 469–476.

(23) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. Accounts Chem. Res. 2018, 51, 1281–1289.

(24) Feng, F.; Lai, L.; Pei, J. Computational chemical synthesis analysis and pathway design. Front. Chem. 2018, 6.

(25) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. J. Chem. Inf. Model. 2011, 51, 2209–2222.

(26) 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.

(27) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. Adv. Neur. In. 2017, 2017-Decem, 2608–2617.

(28) Lei, T.; Jin, W.; Barzilay, R.; Jaakkola, T. Deriving neural architectures from sequence and graph kernels. ICML 2017 2017, 4, 3181–3190.

(29) Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. Angew. Chem. Int. Edit. 2014, 53, 8108–8112.
(30) Schneider, N.; Stiefl, N.; Landrum, G. A. Whats What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* 2016, *56*, 2336–2346.

(31) 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 Cent. Sci.* 2019, *5*, 1572–1583.

(32) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. 2016, 1–19.

(33) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. ”Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* 2018, *9*, 6091–6098.

(34) 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.

(35) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* 2020, *60*, 47–55.

(36) Duan, H.; Wang, L.; Zhang, C.; Guo, L.; Li, J. Retrosynthesis with attention-based NMT model and chemical analysis of ”wrong” predictions. *RSC Adv.* 2020, *10*, 1371–1378.

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

(38) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* 2020, *11*, 3316–3325.
(39) Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. 2014.

(40) Bahdanau, D.; Cho, K. H.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* 2015, 1–15.

(41) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to sequence learning with neural networks. *Adv. Neur. In.* 2014, 4, 3104–3112.

(42) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* 1997, 9, 1735–1780.

(43) Graves, A. Generating Sequences With Recurrent Neural Networks. 2013.

(44) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neur. In.* 2017, 2017-Decem, 5999–6009.

(45) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis. *Lect. Notes Comput. Sc. (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 2019, 11731 LNCS, 817–830.

(46) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comp. Sci.* 2002, 42, 1273–1280.

(47) Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 2005, 18, 602–610.

(48) Lowe, D. M. Extraction of chemical structures and reactions from the literature. Ph.D. thesis, University of Cambridge, 2012.

(49) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017,
(50) Landrum, G. RDKit: Open-Source Cheminformatics Software. 2016; https://github.com/rdkit/rdkit/releases/tag/Release_2020_03_1.

(51) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* 1988, 28, 31–36.

(52) James, C. A.; Weininger, D.; Delany, J. D. Daylight Theory Manual. Daylight Chemical Information Systems Inc. 2002; https://daylight.com/dayhtml/doc/theory/index.html.

(53) Blum, L. C.; Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* 2009, 131, 8732–8733.

(54) Arús-Pous, J.; Blaschke, T.; Ulander, S.; Reymond, J. L.; Chen, H.; Engkvist, O. Exploring the GDB-13 chemical space using deep generative models. *J. Cheminformatics* 2019, 11, 1–33.

(55) Luong, M. T.; Pham, H.; Manning, C. D. Effective approaches to attention-based neural machine translation. *Conf. Proc. - EMNLP 2015 Conf. Empir. Methods Nat. Lang. Process.* 2015, 1412–1421.

(56) Bottou, L. Stochastic Gradient Learning in Neural Networks. *Proceedings of Neuro-Nîmes 1991, 91*, 12.

(57) Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. *ICML 2013 2013*, 2347–2355.

(58) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. In *Adv. Neur. In. 32*; Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.
(59) Riniker, S.; Landrum, G. A. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminformatics* 2013, 5, 1–7.

(60) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* 2015, 7, 1–13.

(61) Guo, Z.; Wu, S.; Ohno, M.; Yoshida, R. A Bayesian algorithm for retrosynthesis. 2020,

(62) 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.

(63) O’Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminformatics* 2016, 8, 1–14.

(64) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Ann. Rep. Comp. Chem.;* Elsevier B.V., 2008; Vol. 4; pp 217–241.

(65) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* 2010, 50, 742–754.

(66) Schomburg, K.; Ehrlich, H.-C.; Stierand, K.; Rarey, M. Chemical pattern visualization in 2D – the SMARTSviewer. *J. Cheminform.* 2011, 3, O12.
Supporting Information Available

MACCS keys assignments.

The set contains the assignments of letters to MACCS keys and list of used keys is presented below.

\{1: 0, 2: 0, 3: 0, 4: 0, 5: 0, 6: 0, 7: 0, 8: 'Fz', 9: 0, 10: 0, 11: 'Mz', 12: 0, 13: 'Pz', 14: 0, 15: 0, 16: 'Bz', 17: 'Wz', 18: 0, 19: 'Cz', 20: 0, 21: 0, 22: 'Rz', 23: 0, 24: 'pz', 25: 'Lz', 26: 'Vz', 27: 0, 28: 0, 29: 0, 30: 0, 31: 0, 32: 'Iz', 33: 'Oz', 34: 'Uz', 35: 0, 36: 'gz', 37: 'Sz', 38: 'tz', 39: 0, 40: 0, 41: 'Tz', 42: 0, 43: 'Dz', 44: 0, 45: 'Gz', 46: 0, 47: 'vz', 48: 0, 49: 0, 50: 'Hz', 51: 'yz', 52: 'nz', 53: 'Ez', 54: 'bz', 55: 'Fz', 56: 0, 57: 'oz', 58: 'mz', 59: 0, 60: 'uz', 61: 'cz', 62: 'yx', 63: 0, 64: 0, 65: 'X', 66: 'Cz', 67: 'rz', 68: 'Yz', 69: 'wz', 70: 0, 71: 'lz', 72: 'iz', 73: 'yz', 74: 'Rx', 75: 'wx', 76: 'Az', 77: 'A', 78: 'Nz', 79: 'vx', 80: 'px', 81: 'Bx', 82: 'Yx', 83: 'F', 84: 'az', 85: 'mx', 86: 'Tx', 87: 0, 88: 'Mx', 89: 'Px', 90: 'Ux', 91: 'Dx', 92: 'bx', 93: 'cx', 94: 'Ox', 95: 'gx', 96: 'M', 97: 'rx', 98: 'R', 99: 'dz', 100: 'V', 101: 'sx', 102: 'Lx', 103: 0, 104: 'Sx', 105: 'ex', 106: 'ox', 107: 0, 108: 'Fx', 109: 'fx', 110: 'G', 111: 'W', 112: 'P', 113: 'dx', 114: 'Vx', 115: 'ez', 116: 'Gx', 117: 'U', 118: 'C', 119: 'sz', 120: 'A', 121: 'm', 122: 'S', 123: 'Y', 124: 'ix', 125: 0, 126: 'E', 127: 'D', 128: 'lx', 129: 'tx', 130: 'Wx', 131: 'Nx', 132: 'hx', 133: 'B', 134: 0, 135: 'nx', 136: 'ux', 137: 'd', 138: 'ax', 139: 'Ex', 140: 'Hx', 141: 'Ix', 142: 'g', 143: 'L', 144: 'p', 145: 'b', 146: 'O', 147: 'T', 148: 'f', 149: 'H', 150: 'y', 151: 'T', 152: 'v', 153: 'u', 154: 'c', 155: 'w', 156: 'n', 157: 'r', 158: 'l', 159: 'h', 160: 's', 161: 'i', 162: 'o', 163: 'a', 164: 't', 165: 'e', 166: 0\}
Figure 4: Distribution of length of product-reactant pairs in Lowe’s USPTO dataset.

Table 6: Hyper-parameter space and hyper-parameters for the best model.

| Parameter          | Possible Values | Best Model Parameters            |
|--------------------|-----------------|----------------------------------|
| RNN Cell Type      | LSTM or Bi-LSTM | Bi-LSTM (Encoder & Decoder)      |
| Number of Layers   | 2, 4, or 6      | 2                                |
| Number of units    | 500,1000, 2000  | 2000                             |
| Learning Rate      | 0.1 - 8         | 4                                |
| Decay factor       | 0.50 - 0.90     | 0.85                             |
| Dropout            | 0.1 - 0.5       | 0.1                              |
| Type of Attention  |                 | Luong’s global attention mechanism\textsuperscript{55} |
| Ground truth | Prediction | $T_r$ | $T_e$ |
|--------------|------------|-------|-------|
| ![Chemical Structure](image1) $\rightarrow$ ![Chemical Structure](image2) + ![Chemical Structure](image3) | ![Chemical Structure](image1) $\Rightarrow$ ![Chemical Structure](image2) + ![Chemical Structure](image3) | 0.82  | 0.92  |
| ![Chemical Structure](image4) $\rightarrow$ ![Chemical Structure](image5) + ![Chemical Structure](image6) | ![Chemical Structure](image4) $\Rightarrow$ ![Chemical Structure](image5) + ![Chemical Structure](image6) | 0.94  | 0.67  |
| ![Chemical Structure](image7) $\rightarrow$ ![Chemical Structure](image8) + ![Chemical Structure](image9) | ![Chemical Structure](image7) $\Rightarrow$ ![Chemical Structure](image8) + ![Chemical Structure](image9) | 0.90  | 0.78  |
| ![Chemical Structure](image10) $\rightarrow$ ![Chemical Structure](image11) + ![Chemical Structure](image12) | ![Chemical Structure](image10) $\Rightarrow$ ![Chemical Structure](image11) + ![Chemical Structure](image12) | 0.97  | 0.76  |
| ![Chemical Structure](image13) $\rightarrow$ ![Chemical Structure](image14) + ![Chemical Structure](image15) | ![Chemical Structure](image13) $\Rightarrow$ ![Chemical Structure](image14) + ![Chemical Structure](image15) | 0.90  | 0.82  |

(a) First set of reactions 1 to 5.
(b) Second set of reactions 6 to 10.

Figure 5: Ten reactions lie in the bioactively similar region used to assess the quality of retrosynthesis.
Table 7: The quantitative summary of the assessment of the specific set containing ten reactions where the candidate reactants lie in the bioactively similar region. Individual scores and sources of errors are stated inside the parenthesis. The FGI or bond disconnection and reactive functional group columns represent the correctness in a True(1)/False(0) fashion. The core structure column presents the averaged accuracy of the core structures of candidate molecules by capturing the correctness of core structures themselves as well as the type and positions of side-substituents. The source of errors are given inside the parenthesis e.g., "C2=0.33, 2/3 fragments" implies that the accuracy of candidate reactant 2 is 0.33 since 2 out of 3 fragments are wrongly predicted. The average of the three criteria is given as a separate column. $T_c$ column represents the averaged $T_c$ values of candidate reactants. C1: Candidate 1, C2 : Candidate 2.

| Reaction Number | FGI or Bond Disconnection | Core Structure | Reactive Functional Group | Avg. $T_c$ |
|-----------------|---------------------------|----------------|---------------------------|------------|
| 1               | 1.00                      | 0.98 (C1=1.00; C2=0.95, 6/5 #C in Alkyl) | 1.00         | 0.99 0.87 |
| 2               | 1.00                      | 0.83 (C1=1.00; C2=0.67, 1/3 fragments) | 1.00         | 0.94 0.81 |
| 3               | 1.00                      | 1.00           | 1.00         | 1.00 0.84 |
| 4               | 1.00                      | 0.75 (C1=1.00; C2=0.50, 1/2 fragments) | 1.00         | 0.92 0.87 |
| 5               | 1.00                      | 0.79 (C1=0.75, 1/2 fragment’s position; C2=0.83, 1/3 fragment’s position) | 1.00         | 0.93 0.86 |
| 6               | 1.00                      | 0.88 (C1=0.75, 1/2 fragment’s position; C2=1.00) | 1.00         | 0.96 0.94 |
| 7               | 1.00                      | 0.96 (C1=0.97, 5/6 #C in ring; C2=0.94, 5/4 #C in Alkyl) | 1.00         | 0.99 0.91 |
| 8               | 1.00                      | 1.00           | 1.00         | 1.00 0.87 |
| 9               | 1.00                      | 1.00           | 1.00         | 1.00 0.83 |
| 10              | 1.00                      | 0.97 (C1=1.00; C2=0.94, position of side subst.) | 1.00         | 0.99 0.85 |
Figure 6: Molecules that are represented with only one bit in the double reactant dataset are given as SMILES strings. SMARTS patterns are visualized by using SMARTSviewer.66