

# Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks

Shuangjia Zheng,<sup>†,‡,⊥</sup> Jiahua Rao,<sup>‡,⊥</sup> Zhongyue Zhang,<sup>‡</sup> Jun Xu,<sup>\*,†,§</sup> and Yuedong Yang<sup>\*,†,||</sup>

<sup>†</sup>Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-sen University, 132 East Circle at University City, Guangzhou 510006, China

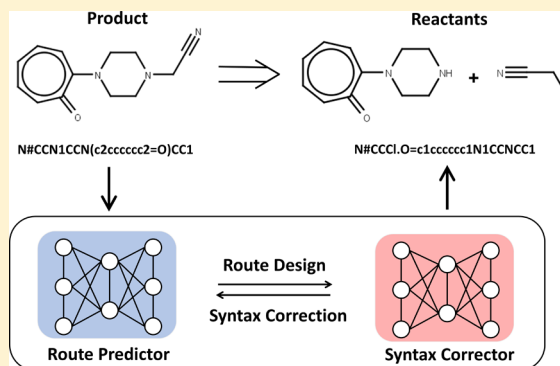
<sup>‡</sup>School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China

<sup>§</sup>School of Computer Science & Technology, Wuyi University, 99 Yingbin Road, Jiangmen 529020, China

<sup>||</sup>Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Ministry of Education, Guangzhou 510000, China

## Supporting Information

**ABSTRACT:** Synthesis planning is the process of recursively decomposing target molecules into available precursors. Computer-aided retrosynthesis can potentially assist chemists in designing synthetic routes; however, at present, it is cumbersome and cannot provide satisfactory results. In this study, we have developed a template-free self-corrected retrosynthesis predictor (SCROP) to predict retrosynthesis using transformer neural networks. In the method, the retrosynthesis planning was converted to a machine translation problem from the products to molecular linear notations of the reactants. By coupling with a neural network-based syntax corrector, our method achieved an accuracy of 59.0% on a standard benchmark data set, which outperformed other deep learning methods by >21% and template-based methods by >6%. More importantly, our method was 1.7 times more accurate than other state-of-the-art methods for compounds not appearing in the training set.



## INTRODUCTION

Organic synthesis is one of the fundamental pillars of modern chemical society, as it provides a wide range of compounds from medicines to materials. The synthetic route to a desired organic compound is commonly constructed by recursively decomposing it into a set of available reaction building blocks.<sup>1</sup> This analysis mode was formalized as a retrosynthesis by Corey,<sup>2,3</sup> who ultimately won the Nobel Prize in 1990.<sup>4</sup>

Planning synthesis requires chemists to accurately predict the disconnections for target molecules. Since molecules may have many possible ways to decompose, the retrosynthetic analysis of a target compound usually leads to a large number of possible synthetic routes. It is challenging to select an appropriate synthesis route because the differences between routes are subtle and often depend on the global structures. Therefore, it remains challenging even for the best chemists to plan a retrosynthetic route for a complex molecule.<sup>5,6</sup>

To this end, many in silico methods have been developed to assist in designing synthetic routes for novel molecules, among which most are dependent on hand-coded reaction templates.<sup>7–15</sup> One of the most famously commercialized computer-assisted software packages is Chematica.<sup>10,16</sup> It contains about 70 000 manually curated reaction transformation rules, which took the Grzybowski Scientific

Invention team more than 15 years. Based on these rules, the synthesis routes can be built according to generalized reaction templates. This kind of method could provide reasonable results when the rule databases include appropriate reaction rules and templates. However, it is not practical to manually encode all the synthesis routes considering the exponential growth in the number of published reactions.<sup>17</sup> Besides, a simple template is generally not enough to reliably predict reactions because it only identifies reaction centers and their neighboring atoms without considering the global information of the target molecule.

Recently, the so-called focused template-based methods were designed through automatically extracting templates from the reaction databases and choosing appropriate rules for the selected templates. The key process of these approaches is to select relevant templates for target molecules. Segler and Waller evaluated template relevance based on molecular fingerprints using artificial neural networks.<sup>18,19</sup> Later, they showed that the Monte Carlo tree combined with deep neural networks could prioritize templates and preselect the most promising retrosynthetic steps.<sup>5</sup> Coley and co-workers also

Received: October 11, 2019

Published: December 11, 2019

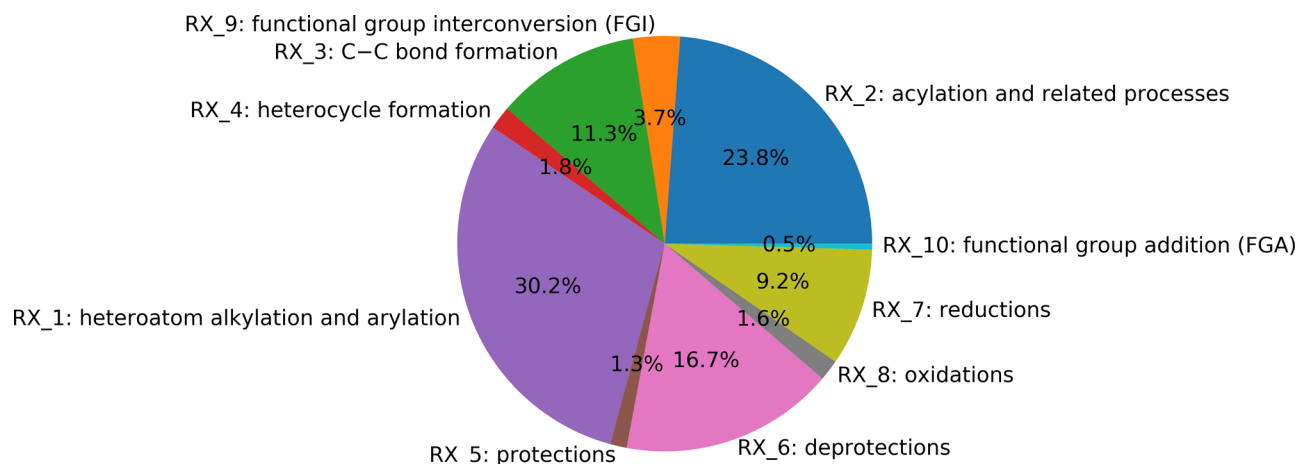


Figure 1. Distribution of reaction classes within the USPTO-50K.

demonstrated an approach for automated retrosynthesis based on the analogy to known reactions.<sup>20</sup> Albeit sturdy, these methods rely heavily on a predefined atom dictionary to map atoms between reactants and products, which is still a nontrivial problem.<sup>21,22</sup> Furthermore, commonly used tools to identify atom mapping are based on databases of expert rules and templates, which seem to get stuck in an infinite loop.<sup>23</sup> At any rate, template-based methods have the limitation that they cannot infer reactions outside the chemical space covered by the template's libraries and thus work poorly when applied to new target structures and reaction types.<sup>19</sup>

To overcome the problem, template-free alternatives have emerged over recent years. The key idea is to use a text representation for the reactants and products (e.g., SMILES); thus, retrosynthesis prediction is converted as machine translations from one language (reactants) to the other (products). Nam and Kim first described a sequence-to-sequence (seq2seq) neural network model for the forward reaction prediction.<sup>24</sup> Later, Liu and co-workers reported a similar seq2seq model that showed comparable performances to the rule-based expert systems.<sup>25</sup> This seq2seq model can be trained in a fully end-to-end fashion that does not require any atom-mapped reaction cases for training. Unfortunately, the model does not show significant improvement in accuracy compared to the rule-based system but produces a great number of chemically invalid outputs.

More recently, transformer architecture has shown advantages in machine translation.<sup>26</sup> It removes traditional recurrent units and is based entirely on the self-attention mechanism, allowing extraction of both local and global features irrespective of the distances between the input and output sequences. Schwaller and co-workers employed this architecture to predict the products of chemical reactions and reached state-of-the-art results,<sup>27</sup> which inspires us to apply the transformer to the prediction of the retrosynthetic reaction.

In this study, we propose a novel template-free self-corrected retrosynthesis predictor (SCROP) built on the multihead attention transformer architecture. Our approach improved the state-of-the-art in retrosynthesis prediction by achieving top-1 accuracies of 59.0% on the Liu's USPTO-50K data set<sup>25</sup> and 41.5% on single product reactions of another Jin's data set.<sup>28</sup> At the same time, the rates of invalid candidate precursors could decrease from 12.1 to 0.7% by coupling with a novel neural network-based syntax checker. When excluding similar targets from the training set, our method achieved an accuracy

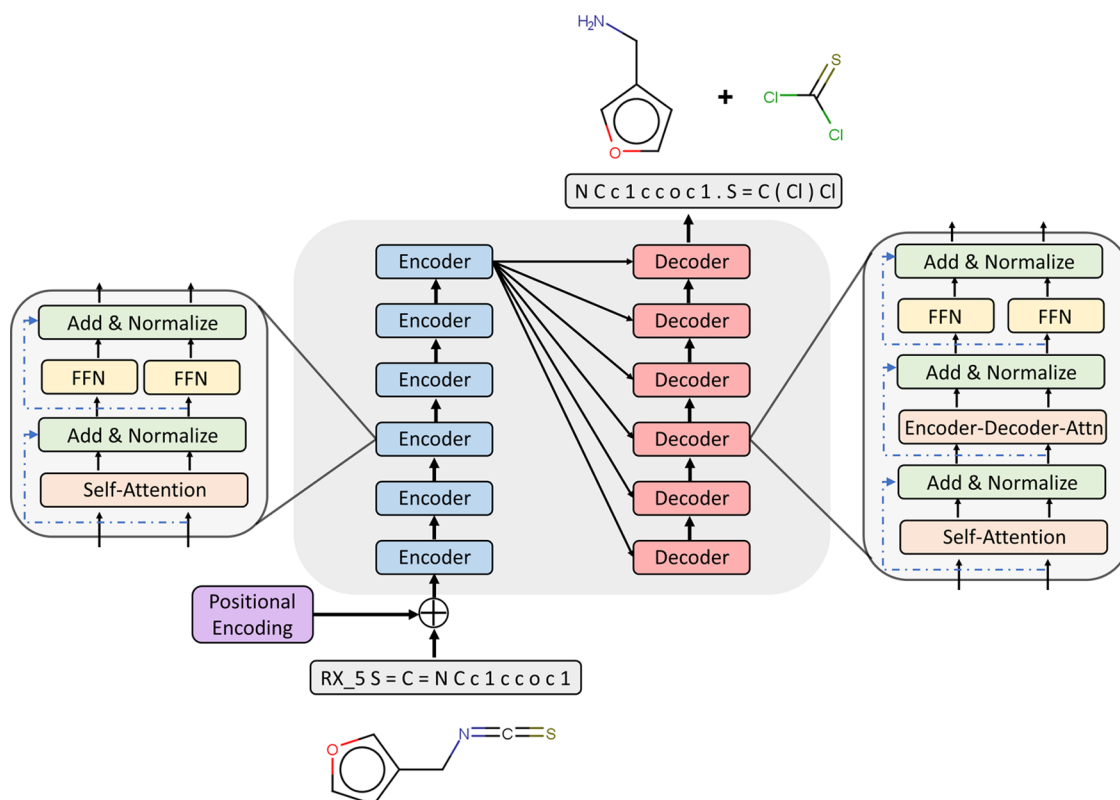
of 47.6%, which was 1.7 times higher than that of other state-of-the-art methods. More importantly, this model requires no handcrafted templates and atom mappings and can accurately predict subtle chemical disconnections.

## OVERVIEW

**Data Sets.** All the reaction data of our experiments were derived from the Lowe's study.<sup>29</sup> For comparison, the first data set we used is a standard benchmark including 50 000 reactions that were later separated into 10 reaction classes by Schneider et al.,<sup>30</sup> namely, USPTO-50K. Figure 1 shows the distribution of each reaction class within the data set. This data set was also employed by Liu et al. and Coley et al. for the same task.<sup>20,25</sup> We followed their random split strategy, with 40K, 5K, and 5K for training, validating, and testing, respectively.

As the random splitting of the data set may cause overestimation by separating similar reactants into the training and test sets, we reconstructed a more challenging data set using a cluster-splitting strategy similar to the one used in the previous study,<sup>31</sup> where reactions were clustered based on topological similarities between target compounds, and reactions belonging to the same cluster were put in the same subset during the splitting of training and test sets. Here, the products were clustered through the typically used Bemis–Murcko atomic frameworks.<sup>32</sup> The similarities of compounds were assessed through the Tanimoto efficient of two-dimensional (2D) fingerprints, and two compounds were considered similar if their similarity is above 0.6. The same proportions (80%/10%/10%) were used to split the reactions into training, validation, and test sets, namely, the USPTO-50K cluster. The splitting makes the retrosynthesis prediction task significantly harder, as the model has to determine the reaction center for a target molecule outside its training (with a low similarity).

In addition, to further evaluate the generalization of our model, we also considered a much bigger data set published by Jin et al (referred to Jin's USPTO). It contains 480 thousand atom-mapped reactions without predefined reaction classes. Following the preprocessing strategies by Liu et al. and Schwaller et al.,<sup>25,33</sup> we removed reagents from reactants and canonicalized the molecules. This data set was divided into 400, 40, and 40K for training, development, and testing purposes, respectively. It should be noted that there are no



**Figure 2.** Overview of the architecture and training procedure of the transformer-based retrosynthetic reaction predictor. “RX\_5” token indicates that the target should be decomposed through a protection reaction.

classified reaction classes in this data set, and thus it is more difficult to predict.

**Problem Definition.** Given an input of a target molecule and its specified reaction type, our task is to predict likely reactants that can react in a specified reaction type to form the target product. In this study, a reaction was described by a variable-length string containing one pair of SMILES notations representing the reactants and target compound. Following the process of Liu et al.,<sup>25</sup> each reaction was split into a source sequence and target sequence for model training. The source sequence is the product of the reaction with a reaction type token prepended to the sequence, and the target sequence is the reactant set. For example, a protection reaction can be described as “NCc1ccoc1.S=(Cl)Cl>>[RX\_5]S=C=NCc1ccoc1”, where “S=C=NCc1ccoc1” is the source sequence, “NCc1ccoc1.S=(Cl)Cl” is the target sequence, and “[RX\_5]” is the reaction class 5 (protections).

These sequences are then encoded into one-hot matrices with a token vocabulary (in our case, it has a total of 50 unique tokens retrieved from the data set). In the one-hot encoding approach, each sequence is represented by a set of token vectors. All token vectors have the same number of components. Each component in a vector is set to zero except the one at the token’s index position. To make the training procedure more stable and efficient, the input one-hot matrices are compressed to information-enriched word-embedding vectors following the previous work.<sup>34,35</sup> As a result, each input sequence is finally represented in a molecular embedding

$$m = (t_1, t_2, \dots, t_n) \quad (1)$$

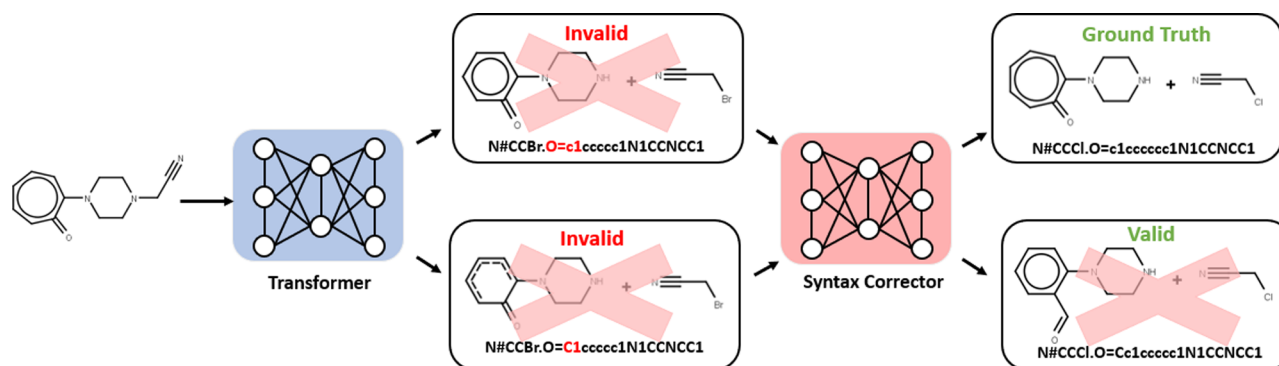
where  $t_i$  is a vector standing for a  $d$ -dimensional token embedding for the  $i$ th token in a molecule consisting of  $n$  tokens.

**Model.** Our model predicts the synthetic route for a target molecule in a two-step manner: (1) applying a fully trained retrosynthetic reaction predictor to infer a set of raw candidate reactants and (2) fixing their syntax errors to make more reasonable predictions using a molecular syntax corrector.

**Retrosynthesis Predictor.** In the first stage, we adapted the transformer architecture to map the sequence of the products to the sequence of the reactants.<sup>26</sup> As shown in Figure 2, the architecture of the transformer system follows the so-called encoder–decoder paradigm to be trained in an end-to-end fashion. The encoder layers are input with the source molecular embedding  $m_s = (t_1, \dots, t_n)$  and iteratively transform it into a latent representation  $l = (l_1, \dots, l_n)$ . After finishing the encoder phasing, each step in the decoding phase outputs a token based on the latent information  $l$  until the ending token “/s” is reached to indicate the completion of output by the transformer decoder. The predicted output  $y_p = (y_1, \dots, y_m)$  is then used to compare with target reactants sequence  $m_r = (t_1, \dots, t_k)$ , and the training goal is to minimize the gap between  $y_p$  and  $m_r$  so that the model can finally infer accurate reactions.

Several identical layers are stacked for the encoding phase. Each layer comprises a combination of a multihead self-attention sublayer and a positional feedforward network (FFN) sublayer. A residual connection and layer normalization were employed to integrate the two sublayers.<sup>36</sup>

Different from the encoder, the decoder is composed of two types of attention multihead attention layers: (i) a decoder self-attention and (ii) an encoder–decoder attention. The decoder self-attention focuses on the previous predictions of



**Figure 3.** Example SMILES syntax correction for two invalid predictions generated by transformer-based retrosynthesis predictor. The syntax corrector fixes the syntax errors and produces the ground-truth reactants.

reactants made step by step, masked by one position. The encoder–decoder attention builds the connection between the final encoder representation and the decoder representation. It integrates the information of the source molecular embeddings with the reactant strings that have been predicted so far and thus helps the decoder to focus on appropriate places in the input sequence.

A multihead attention unit itself comprised several scaled-dot attention layers performing the attention mechanism in parallel, which are then concatenated and projected to the final values. The scaled-dot attention layers take three matrices: the queries  $Q$ , the keys  $K$ , and the values  $V$ . The query, key, and value matrices are created by multiplying the input molecular embedding  $M$  by three weight matrices that were also trained during the training process. We then compute the attention weight for each token within a SMILES string as follows

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2)$$

where  $d_k$  is a scaling factor depending on the size of weight matrices. The dot-product of the keys and the queries computes how closely the keys are correlated with the queries. The dot-product is large if the query and key are aligned well. Each key has a corresponding value vector, which is multiplied with the output of the softmax. By this procedure, the encoder extracts pivotal features from the source sequence, which are then queried by the decoder depending on its preceding outputs. Thus, the model can learn the global-level information from the input molecular embeddings and build a semantic connection between the encoder and decoder.

As the transformer architecture removes the recurrent units from traditional recurrent networks, the model lacks a way to account for the order of words in the input SMILES strings. To address this, we used the position encoding as proposed in the previous study,<sup>26</sup> which adopted the sine and cosine functions to identify the position of different tokens in the sequence

$$\begin{aligned} PE_{(\text{pos}, 2i)} &= \sin\left(\frac{\text{pos}}{\text{timescale}^{2i/d_{\text{emb}}}}\right) = PE_{(\text{pos}, 2i+1)} \\ &= \cos\left(\frac{\text{pos}}{\text{timescale}^{2i/d_{\text{emb}}}}\right) \end{aligned} \quad (3)$$

where  $\text{pos}$  is the position and  $i$  is the dimensional index of position encodings. The outputs of positional encodings have the same dimension  $d_{\text{emb}}$  as the token embeddings. The

timescale is set to 10 000 to form a geometric progression from  $2\pi$  to  $10\,000 \times 2\pi$ .

For a particular source sequence, the training objective is to minimize the cross-entropy loss function

$$\mathcal{L}(y, m) = - \sum_{i=1}^K y_i \log(m_i) \quad (4)$$

where  $m_i$  and  $y_i$  are the predicted and actual values at the  $i$ th position for the target molecular sequences, respectively.

Our best-performance models were trained for 12 h on one GPU (Nvidia 1080TI) on the training set, saving one checkpoint every 20 000 steps and averaging the last ten checkpoints. More detailed hyperparameter settings of the model are shown in Table S1. The hyperparameters were chosen according to the performances on the validating set. A beam search procedure<sup>37</sup> was then used to infer multiple reactant candidates on the test set. We used the best-performance model to infer the reactant candidates with a beam width of 10. Therefore, the top ten candidate sequences ranked by overall probability are retained.

**Molecular Syntax Corrector.** It is important to note that syntactically plausible molecular strings are not guaranteed to be semantically valid. For instance, “c1ccoc” cannot be deduced to a valid structure because it misses the token “1” representing the closing of the heterocycle. Previous works entirely relied on the raw outputs obtained from the default beam search,<sup>25,33</sup> which enumerates the top  $N$  predictions based on the joint probability of generated tokens without considering the chemical feasibility. Here, we further built a transformer-based syntax corrector to automatically correct the syntax of unreasonable SMILES strings for improving the performances. The design of the neural network was motivated by the grammatical error correction tool widely utilized in natural language processing tasks.<sup>38</sup> Figure 3 illustrates the procedure of our syntax correction system. The syntax corrector takes the unreasonable predictions generated from the retrosynthetic reaction predictor and fixes their syntax errors to increase the quality of predictions.

The system does this by taking ground-truth reactants and invalid reactants generated from the retrosynthesis predictor to produce input–output pairs (where the output is the ground-truth reactants), which are then used to train a sequence-to-sequence transformer model. Concretely, we first use a fully trained model to generate the top ten candidate precursors given in a set of target compounds in the training set. Second, we filter the candidate reactant sets by removing the ground



truth-reactant sets corresponding to the target molecules. Third, we construct a training library that consists of a set of input–output pairs, where the inputs are predicted invalid reactants, and the outputs are the ground-truth reactants. Given such a syntax correction data set with input–output pairs, we can train a new sequence-to-sequence model using the transformer architecture introduced above and hook it up to the retrosynthetic reaction predictor. Once trained, we can input unsatisfactory SMILES strings generated from the reaction predictor and fix their syntax errors to make more reasonable predictions. Note that we only retained the top-1 candidate produced by the syntax corrector and replace the original one.

After correction, we canonicalize all predicted sequences by reordering the tokens with fixed rules and compared the predicted candidates with the ground-truth reactants. All scripts were written in Python (version 3.5), and RDKit was used for data preprocessing and molecule canonicalization.<sup>39</sup> The transformer model was implemented using OpenNMT.<sup>40</sup>

## EXPERIMENTS AND RESULTS

We evaluated the self-corrected retrosynthesis predictor (SCROP) on USPTO-50K data sets with two split methods (random and clustering) and compared the performance with other state-of-the-art results (we have repeated their results as reported in their original papers). Table 1 shows the

**Table 1. Comparison of Top-N Accuracies between the Baselines and SCROP on USPTO-50K**

| data                   | model        | top-n accuracy (%), n = |             |             |             |
|------------------------|--------------|-------------------------|-------------|-------------|-------------|
|                        |              | 1                       | 3           | 5           | 10          |
| with reaction class    | Liu-baseline | 35.4                    | 52.3        | 59.1        | 65.1        |
|                        | Liu-seq2seq  | 37.4                    | 52.4        | 57.0        | 61.7        |
|                        | similarity   | 52.9                    | 73.8        | <b>81.2</b> | <b>88.1</b> |
|                        | SCROP-noSC   | 58.8                    | 74.4        | 77.5        | 80.1        |
|                        | SCROP        | <b>59.0</b>             | <b>74.8</b> | 78.1        | 81.1        |
| without reaction class | similarity   | 37.3                    | 54.7        | 63.3        | 74.1        |
|                        | SCROP-noSC   | 43.3                    | 59.1        | 64.0        | 67.0        |
|                        | SCROP        | <b>43.7</b>             | <b>60.0</b> | <b>65.2</b> | 68.7        |

The bold values are the highest ones among the corresponding columns.

quantitative performance of the models on the USPTO-50K data set when the reaction type is known and unknown, respectively. When inferring reaction within a specific reaction class, the SCROP outperforms all baselines in the top-1 recommendation, achieving an exactly matching accuracy of 59.0%. The percentages of correctly predicted reactants by top-3, top-5, and top-10 are 74.8, 78.1, and 81.1%, respectively. These results are much higher than those reported in the previous seq2seq model (37.4% for top-1 and 61.7% for top-10).

Note that the template-based model of Coley et al. (similarity) predicted 100 candidates and picked the top-10 as a result. As the SCROP did not rank candidates but was trained for accurately predicting the top-1 outcome and only predicted ten candidates, it is not surprising that the similarity-based method has higher top-5 and top-10 accuracies. Even so, the SCROP improves the similarity-based method by a margin of 6.1% in the top-1 accuracy.

A comparison with the results of the original retrosynthesis predictor (SCROP-noSC) showed that the syntax corrector

leads to consistent increases of top-1 to top-10 accuracies, ranging from 0.2 to 1.0%. The growth of top-1 and top-3 accuracies is not apparent because only a small part of invalid predictions was generated in these two stages.

Without prior knowledge of the reaction class (removing the reaction type tokens in the training procedure), the SCROP still improves over the similarity-based method by a margin of 6.4 and 5.7% for top-1 and top-3 accuracies and achieves a comparable result in the top-5 and top-10 suggestions.

**Generalization Estimation.** To compare the generalization ability of these approaches, we further evaluated the models on the USPTO-50K cluster data set. We retrained both the seq2seq and similarity-based methods on the USPTO-50K cluster with the same parameter settings as in the original papers, except that the generated candidates in the similarity-based method were set to 10 rather than 100 for a fair comparison. As shown in Table 2, the SCROP achieved an

**Table 2. Comparison of Top-N Accuracies between the Baselines and SCROP on USPTO-50K Cluster**

| data                   | model      | top-n accuracy (%), n = |             |             |             |
|------------------------|------------|-------------------------|-------------|-------------|-------------|
|                        |            | 1                       | 3           | 5           | 10          |
| with reaction class    | seq2seq    | 25.5                    | 38.7        | 43.6        | 49.0        |
|                        | similarity | 36.7                    | 58.0        | 61.4        | 67.2        |
|                        | SCROP      | <b>47.6</b>             | <b>63.9</b> | <b>68.1</b> | <b>71.1</b> |
| without reaction class | seq2seq    | 16.5                    | 28.8        | 34.0        | 40.6        |
|                        | similarity | 20.7                    | 36.3        | 43.2        | 46.9        |
|                        | SCROP      | <b>36.2</b>             | <b>52.0</b> | <b>57.1</b> | <b>60.9</b> |

The bold values are the highest ones among the corresponding columns.

order-of-magnitude improvement over baselines with or without the reaction class from top-1 to top-10 predictions. Besides, we found that our model performed more stably compared to the template-based method: the top-1 accuracy of our model decreases only by 7.5% compared with 16.6% of the similarity method when planning retrosynthesis for clustered data set without the knowledge of reaction class. This result demonstrates that our template-free method has a better generalization ability than the template-based one when the training data set has no similar compound to the target compounds.

We also computed top-10 results for each reaction class on the USPTO-50K cluster data set. As shown in Table 3, SCROP consistently performs better than the seq2seq model in all 10 reaction classes and betters the similarity model in 6 reaction classes. For example, the 4th reaction class (heterocycle formation) commonly includes the formation of cyclic structures, resulting in a significant difference between the reactant SMILES string and the target product SMILES string. The best performance by SCROP indicates its ability to induce better syntactic relationships and capture global chemical information from the reaction data. The SCROP also outperforms the seq2seq and similarity models by a large margin in the 1st class (heteroatom alkylation and arylation). The key feature of this reaction class is that the reactions possibly occur with many different functional groups within the target molecules. It is hard to identify the accurate reaction site when the structure of target molecules is not similar to the one in the knowledge base. As a result, the SCROP shows better generalization ability in this reaction type and infers reactants correctly.

Table 3. Model Top-10 Accuracy within Each Class on USPTO-50K Cluster When the Reaction Type is Known

| model      | top-10 accuracy (%), reaction class = |             |             |             |             |      |             |             |             |             |
|------------|---------------------------------------|-------------|-------------|-------------|-------------|------|-------------|-------------|-------------|-------------|
|            | 1                                     | 2           | 3           | 4           | 5           | 6    | 7           | 8           | 9           | 10          |
| seq2seq    | 43.9                                  | 58.7        | 29.8        | 11.9        | 63.5        | 54.4 | 66.0        | 54.4        | 43.3        | 45.8        |
| similarity | 56.4                                  | 77.0        | 50.8        | 52.4        | <b>86.2</b> | 77.5 | 73.6        | <b>86.2</b> | 58.8        | <b>85.3</b> |
| SCROP      | <b>71.8</b>                           | <b>78.9</b> | <b>57.2</b> | <b>56.2</b> | 85.1        | 68.5 | <b>79.9</b> | 69.6        | <b>68.0</b> | 68.8        |

The bold values are the highest ones among the corresponding columns.

Table 4 shows a comparison with the seq2seq model from Liu et al. on the Jin's USPTO data set. By achieving 41.5 and

Table 4. Top-1, -3, -5, and -10 Accuracies of Different Models on the Jin's USPTO Data Set

| models  | top- <i>n</i> accuracy (%) |             |             |             |
|---------|----------------------------|-------------|-------------|-------------|
|         | 1                          | 3           | 5           | 10          |
| seq2seq | 21.3                       | 33.1        | 37.0        | 40.6        |
| SCROP   | <b>41.5</b>                | <b>53.3</b> | <b>56.7</b> | <b>59.2</b> |

The bold values are the highest ones among the corresponding columns.

59.2%, respectively, for the top-1 and top-10 accuracies, SCROP outperforms the state-of-the-art template-free model in a larger chemical space, demonstrating its scalability. We note that the similarity method developed by Coley et al. achieves a top-1 accuracy of 5.1% and top-10 10.2% on the Jin's USPTO test set with the pretrained model on the USPTO-50K. The similarity method was not retrained on Jin's USPTO training set because the unreliable atom mappings in this data set led to many program bugs, also indicating the limitation of template-based methods.

**Evaluation of Syntax Corrector.** In Liu's work,<sup>7</sup> incorrect predictions were summarized as of three types: grammatically invalid outputs, grammatically valid but chemically unreasonable, and chemically plausible but not matching to the ground truth. In this study, we only used the syntax corrector to fix grammatically invalid outputs since the identification of the other two types of errors requires the knowledge of ground truth reactants. As shown in Figure 4, for SCROP, only 0.7% of the top-1 and 2.3% of the top-10 predictions are grammatically invalid, which is significantly better than those by the retrosynthesis predictor (SCROP-noSC) and the seq2seq model. As an example, Figure 5 shows the predicted synthesis route of Benoxaprofen, an anti-inflammatory drug, coupling with the syntax corrector. The model successfully proposes the

ground truth by correcting the original rank 2 prediction, which is an invalid molecule string. Additionally, the syntax corrector makes the 3<sup>rd</sup> ranked prediction more reasonable, although it is not the correct answer. These results suggest that the syntax corrector can be used to optimize the invalid predictions. It provides a new strategy to optimize the synthesis route when the predicted reactions are proven wrong. More examples of accurately corrected reactions are shown in the Supporting Information.

**Attention Analysis.** To investigate what the model has learned, we further analyzed the attention weights in our model. The attention weights provide clues on tokens in the input string that were considered to be more critical when a particular symbol in the output string was generated. Figure 6, for example, shows the top-1 candidate's attention maps of an acylation reaction extracted from the SCROP (accurately predicted) and seq2seq (wrong predicted) models. We observed that strong weights trend diagonally in SCROP's attention map, which constantly aligns with SMILES substrings that are shared between the input and output. Besides, when inferring the unseen reaction sites "OH" and ".", the model can simultaneously take several noncontinuous tokens into account. This suggests that the SCROP tries to extract both the local and global information of the source sequence. In comparison, the attention map of the seq2seq model fails to align the substrings of the input–output pair and results in a wrong prediction.

**Limitation.** A distinct disadvantage is that the model scores candidates by only taking into account the overall probability of the predicted strings. Practically, there are many additional considerations in synthetic route design, such as process complexity, expense, productivity, safety, and so on. This is because the public data sets do not contain additional technical information except for the reaction strings itself. We would explore the scoring system by incorporating additional considerations like similarity, reaction complexity, cost, and reaction yield to optimize the present evaluation metric.

Another limitation of the model is the multistep reaction. A possible option would be to recursively decompose the target compound using the transformer prediction system until the commercially available building block is obtained. Monte Carlo tree search could be employed to score the outputs in each single reaction step.<sup>5</sup>

Besides, finding the best-performing set of hyperparameters for a deep neural network as well as the inference procedure is computationally expensive, as many hyperparameter settings take tens of hours to train using one GeForce GTX1080Ti graphics card. For this reason, we use a rough grid search to identify the final settings in training and inferring procedure. Better optimization techniques with sufficient equipment may result in a further increase in accuracy.

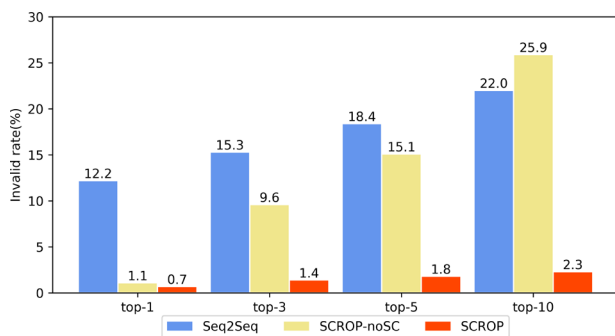
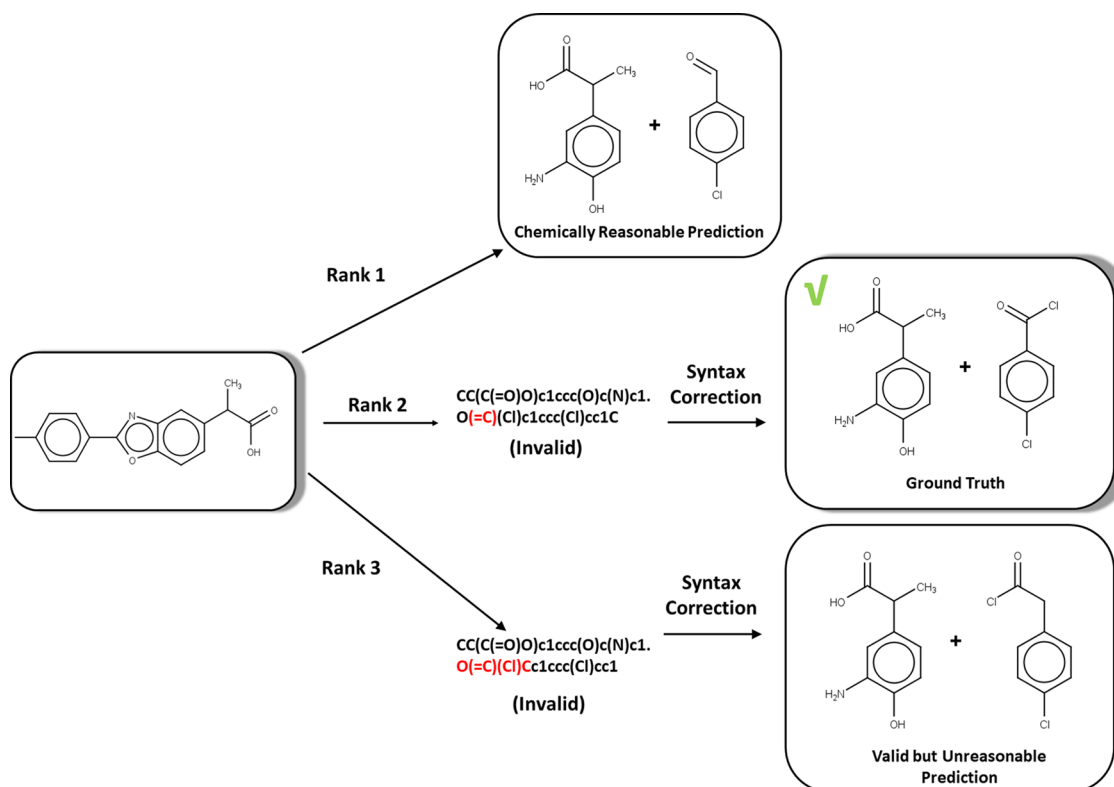
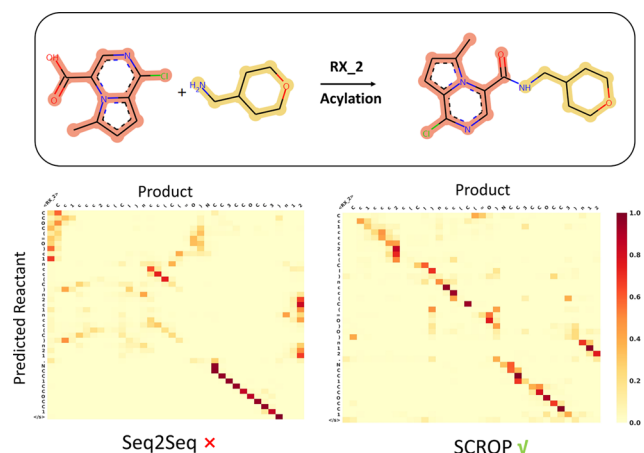


Figure 4. Comparison of invalid rates among seq2seq, retrosynthesis predictor (SCROP-noSC), and self-corrected retrosynthesis predictor (SCROP) for different beam sizes.



**Figure 5.** Example of retrosynthetic predictions coupling with the syntax corrector. The model successfully proposes the ground truth by correcting the original rank 2 prediction, which is an invalid molecule string. The red color denotes incorrect grammar.



**Figure 6.** Top-1 candidate's attention maps of an acylation reaction extracted from the SCROP (accurately predicted) and seq2seq (wrong predicted) models.

## CONCLUSIONS

In this work, we have proposed a novel template-free deep learning method for chemical retrosynthetic prediction. Instead of generating candidate precursors by reaction templates, we employed transformer neural networks to generate potential candidates by learning the sequential representation of the reactant–product pairs. Our approach improved the state of the art in retrosynthesis prediction by achieving top-1 accuracies of 59.0% on the Liu's USPTO-50K data set and 41.5% on the Jin's data set. At the same time, the rates of invalid candidate precursors decreased from 12.1 to 0.7% by coupling with a novel neural network-based syntax checker. When excluding similar targets from the training set,

our method achieved an accuracy of 47.6%, which is 1.7 times higher than those of other methods. More importantly, our method is trained in an end-to-end fashion and is free of chemical rules, and the accuracy will improve automatically with the increase of the training samples.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.9b00949>.

Key hyperparameters; parameters for the best model in bold; randomly selected accurately corrected examples from the syntax correction system; and example of SMILES syntax correction for two kinds of error predictions (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

\*E-mail: [junxu@biochemomes.com](mailto:junxu@biochemomes.com) (J.X.).

\*E-mail: [yangyd25@mail.sysu.edu.cn](mailto:yangyd25@mail.sysu.edu.cn) (Y.Y.).

### ORCID

Shuangjia Zheng: 0000-0001-9747-4285

Jun Xu: 0000-0002-1075-0337

Yuedong Yang: 0000-0002-6782-2813

### Author Contributions

<sup>†</sup>S.Z. and J.R. contributed equally to this work.

### Author Contributions

S.Z., J.R., and Y.Y. contributed concept and implementation. S.Z. and J.R. co-designed experiments. S.Z. and J.R. were responsible for programming. All of the authors contributed to the interpretation of results. S.Z. and Y.Y. wrote the



manuscript. All of the authors reviewed and approved the final manuscript.

## Notes

The authors declare no competing financial interest.

The project and data will be made available at: <https://github.com/sysu-yanglab/Self-Corrected-Retrosynthetic-Reaction-Predictor>.

## ACKNOWLEDGMENTS

The work was supported in part by the National Key R&D Program of China (2018YFC0910500), GD Frontier & Key Tech Innovation Program (2018B010109006 and 2019B020228001), the National Natural Science Foundation of China (61772566, U1611261, and 81801132), and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2016ZT06D211).

## REFERENCES

- (1) Robinson, R. LXIII.—A synthesis of tropinone. *J. Chem. Soc., Trans.* **1917**, 111, 762–768.
- (2) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, 166, 178–192.
- (3) Corey, E. J. General methods for the construction of complex molecules. *Pure Appl. Chem.* **1967**, 14, 19–38.
- (4) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-assisted analysis in organic synthesis. *Science* **1985**, 228, 408–418.
- (5) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, 555, 604–610.
- (6) Collins, K. D.; Glorius, F. A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* **2013**, 5, 597.
- (7) 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. Comput. Sci.* **1995**, 35, 34–44.
- (8) 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.* **2012**, 51, 7928–7932.
- (9) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* **2012**, 52, 1745–1756.
- (10) 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.
- (11) 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.
- (12) Ihlenfeldt, W. D.; Gasteiger, J. Computer-assisted planning of organic syntheses: the second generation of programs. *Angew. Chem., Int. Ed.* **1996**, 34, 2613–2633.
- (13) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, 34, 247–266.
- (14) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to predict chemical reactions. *J. Chem. Inf. Model.* **2011**, 51, 2209–2222.
- (15) Bøgevig, A.; Federsel, H.-J. r.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route design in the 21st century: The IC SYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* **2015**, 19, 357–368.
- (16) Corey, E.; Long, A. K.; Greene, T. W.; Miller, J. W. Computer-assisted synthetic analysis. Selection of protective groups for multistep organic syntheses. *J. Org. Chem.* **1985**, 50, 1920–1927.
- (17) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, 51, 1281–1289.
- (18) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, 23, 5966–5971.
- (19) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, 23, 6118–6128.
- (20) 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.
- (21) Jaworski, W.; Szymkuc, S.; Mikulak-Klucznik, B.; Piecuch, K.; Klucznik, T.; Kazmierowski, M.; Rydzewski, J.; Gambin, A.; Grzybowski, B. A. Automatic mapping of atoms across both simple and complex chemical reactions. *Nat. Commun.* **2019**, 10, No. 1434.
- (22) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic reaction mapping and reaction center detection. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, 3, 560–593.
- (23) 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.
- (24) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions, arXiv:1612.09529. arXiv.org e-Print archive. <https://arxiv.org/abs/1612.09529> (accessed December 29, 2016).
- (25) 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.
- (26) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need, arXiv:1706.03762. arXiv.org e-Print archive. <https://arxiv.org/abs/1706.03762> (accessed June 12, 2017).
- (27) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. A. Molecular transformer for chemical reaction prediction and uncertainty estimation, arXiv:1811.02633. arXiv.org e-Print archive. <https://arxiv.org/abs/1811.02633> (accessed November 6, 2018).
- (28) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network; Advances in Neural Information Processing Systems, 2017; pp 2607–2616.
- (29) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*; University of Cambridge, 2012.
- (30) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, 56, 2336–2346.
- (31) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 579–586.
- (32) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, 39, 2887–2893.
- (33) Schwaller, P.; Gaudin, T.; Lanyi, 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) Levy, O.; Goldberg, Y. *Neural Word Embedding as Implicit Matrix Factorization*; Advances in Neural Information Processing Systems, 2014; pp 2177–2185.
- (35) Zheng, S.; Yan, X.; Yang, Y.; Xu, J. Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, 59, 914–923.
- (36) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition, arXiv:1512.03385. arXiv.org e-Print archive. <https://arxiv.org/abs/1512.03385> (accessed December 10, 2015).
- (37) Ow, P. S.; Morton, T. E. Filtered beam search in scheduling. *Int. J. Prod. Res.* **1988**, 26, 35–62.
- (38) Ng, H. T.; Wu, S. M.; Briscoe, T.; Hadiwinoto, C.; Susanto, R. H.; Bryant, C. In *The CoNLL-2014 Shared Task on Grammatical Error Correction*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task, 2014; pp 1–14.
- (39) RDKit: Open-source cheminformatics, <http://www.rdkit.org> (accessed September 3, 2018).



(40) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation, arXiv:1701.02810. arXiv.org e-Print archive. <https://arxiv.org/abs/1701.02810> (accessed January 10, 2017).