RetroPrime: A Chemistry-Inspired and Transformer-based Method for Retrosynthesis Predictions

Xiaorui Wang[†], Jiezhong Qiu[‡], Yuquan Li[†], Guangyong Chen[§], Huanxiang Liu[⁎], Benben Liao[*,⫽], Chang-Yu Hsieh[*,⫽], Xiaojun Yao[*,†]

[†]College of chemistry and chemical engineering, Lanzhou University, Lanzhou, China.

[‡]Department of Computer Science and Technology, Tsinghua University, Beijing, China.

[§]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen, Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

[⁎]School of pharmacy, Lanzhou University, Lanzhou, China.

[⫽]Tencent Quantum Laboratory, Tencent, Shenzhen, China.

Corresponding authors:

Xiaojun Yao

E-mail: xjyao@lzu.edu.cn.

Chang-Yu Hsieh

Email: kimhsieh@tencent.com.

Benben Liao

Email: bliao@tencent.com.

ABSTRACT

Retrosynthesis prediction is a crucial task for organic synthesis. In this work, we propose a template-free and Transformer-based method dubbed RetroPrime, integrating chemists' retrosynthetic strategy of (1) decomposing a molecule into synthons then (2) generating reactants by attaching leaving groups. These two steps are accomplished with versatile Transformer models, respectively. While RetroPrime performs competitively against all state-of-the-art models on the standard USPTO-50K dataset, it manifests remarkable generalizability and outperforms the only published result by a non-trivial margin of 4.8% for the Top-1 accuracy on the large-scale USPTO-full dataset. It is known that outputs of Transformer-based retrosynthesis model tend to suffer from insufficient diversity and high invalidity. These problems may limit the potential of Transformer-based methods in real practice, yet no prior works address both issues simultaneously. RetroPrime is designed to tackle these challenges. Finally, we provide convincing results to support the claim that RetroPrime can more effectively generalize across chemical space.

Keywords: Deep Learning, Natural Language Processing, Template-free Single-Step Retrosynthesis.

Introduction

Organic synthesis is not only an essential part of organic chemistry but also a cornerstone for a wide array of modern scientific disciplines such as drug discovery, environmental science, and materials science etc. Retrosynthetic analysis is the most common method to design synthetic routes by iteratively decomposing molecules into potentially simpler and easier-to-synthesize precursors via applying known reactions[1]. In recent years, with the development of artificial intelligence technology, computer-aided synthesis planning (CASP) has further empowered chemists to contemplate even more complex molecules and save tremendous amount of time and energy to design synthetic experiments[2,3,4,5,6,7,8,9,10,11,12,13,14,15].

At present, purely machine-learning models are classified into two categories[16]: the template-based[4,17,18] and template-free[19,20,21,22,23,24] methods. A template-based algorithm extracts reaction templates from chemical data[25,26], matches the subgraph in the product part of the template to a target molecule, decomposes the target molecules as prescribed by the matched template, and completes the leaving group through the atomic changes indicated by the template to obtain the reaction precursors. Despite being interpretable in terms of why certain templates are preferred, template-based methods can only predict reactions if corresponding templates have been curated in a database[4,17]. With ever growing list of reaction templates, it is certainly desirable to contemplate alternative approaches.

It is often claimed that template-free methods may predict chemical reactions not present in a training set. However, this intriguing aspect of template-free methods has only been studied and reported in one reference[31] so far. In the **Supplementary Information** section S3, we present one experiment to support this intuition. In particular, we show template-free methods perform much better in predicting reactions when the corresponding templates never appear in the training set. In this work, we focus on Transformer-based template-free method.

Liu et al.[20] treated the one-step retrosynthesis as a translation task, using SMILES[27] to represent molecules and using an LSTM[28] model, a venerable tool in natural language processing (NLP), to convert SMILES of a product to SMILES of reactant(s). Later on, many researchers[19,29,30,31,32] adopted more advanced NLP model, the Transformer[33], for predicting retrosynthesis. Transformer-based methods easily outperform the baseline established by the prior art. Furthermore, the same model architecture can be directly applied for 'forward prediction'[34], i.e. predicting a product molecule given a set of reactants and reagents. In another study[32] , Lee et. al. unambiguously demonstrated that generalizability of Transformer model across chemical spaces. Transformer not only performs well in single-step retrosynthesis, but also in multi-step retrosynthesis. Lin et al.[19] combined Transformer and Monte-Carlo tree search, and re-discovers the reported retrosynthetic route of four molecules.

While Transformer-based models possess so many desiderata, they suffer from two severe shortcomings: (1) lack of diverse outputs[29], and (2) chemically invalid outputs. So far, these difficulties have not been intensively discussed in the chemistry literature and are partially diverted by the fact that Transformer-based models perform well under the metric of Top-N accuracy. This metrics however is not entirely appropriate for retrosynthesis. Schwaller et al.[35] proposed a multifaceted evaluation scheme to replace the Top-N accuracy that could capture these two subtle issues to some extent. While, in this work, we still stick to Top-N accuracy in order to offer a consistent comparison with other methods reported in the literature, but we also discuss these two shortcomings in depth.

There are only a few studies set out to address either of these two shortcomings. For instance, to reduce the number of grammatically invalid SMILES outputted by a Transformer, Zheng et al.[30] proposed a self-correction learning scheme. While this method reduces the number of invalid SMILES, which can be easily detected, it does not guarantee corrected outputs are necessarily legitimate reactants. In a separate study, Chen et. al.[29] attempted to coax a Transformer into giving more diverse outputs covering a broader set of reactions. This successful demonstrations by Chen et. al. is encouraging, but the overall Top-N accuracy of this model does not reach the state-of-the-art results. Further details on these two shortcomings are elaborated in the **Results and Discussion**.

Herein, we set out to improve upon both shortcomings while achieving the state-of-the-art results. We name our method the RetroPrime. Following a recent trend[21,23] to imitate a chemist's approach to retrosynthesis in two steps: (1) disconnect a molecule at a reaction center, and (2) convert synthons into reactants; RetroPrime relies on two Transformers to predict reaction center and synthons-to-reactants, respectively. This two-step framework simplifies the complex pattern of chemical reactions for Transformer to learn in a divide-and-conquer manner. To enhance output diversity and chemical validity, we introduce the "mix and match" and "align and label" strategies in the RetroPrime workflow. Details may be found in **Methodology**.

We have not only evaluated our methods on a standard dataset USPTO-50K[36] but also tested on the large-scale USPTO-full[4], which is one of few results for template-free methods tested with roughly a million records of reaction data. It is remarkable that RetroPrime enjoys a lead of 4.8% for Top-1 accuracy over the state-of-the-art template-based method GLN[4] when tested on the USPTO-full. Finally, in the **Section Generalizability across chemical space**, we conduct a more detailed experiment to show that RetroPrime exhibits superior generalizability for making predictions across chemical spaces in comparison to another two retrosynthesis algorithms: Sing-Stage Transformer (Abbreviated as S-Transformer) and RetroSim[37].

By substantially improving Transformer's shortcoming while achieving state-of-the-art performances, RetroPrime is a versatile tool and points out a promising direction to further develop more advanced template-free methods that, hopefully, may enable fully automated and data-driven retrosynthetic planning of complex molecules in the future.

Results and Discussion

Bird's-eye view.

Following chemists' approach, we solve a one-step retrosynthesis in two stages. 1. Given a molecule, identify possible reaction centers and disconnect relevant bonds to produce synthons (*P*→*S*). 2. Transform synthons to reactants (*S*→*R*). Both tasks can be accomplished with advanced deep-learning techniques. In particular, we employ the powerful Transformer model, commonly used for the natural language processing, in both steps. **Figure 1** provides a bird's-eye view on our proposed method pipeline.

In this work, we refer to the two Transformers as the product-to-synthons (P2S) model and the synthons-to-reactants (S2R) model, respectively. The workflow is summarized as follows. Firstly, the P2S model tags atoms in a molecule that may potentially participate in a reaction. Multiple possibilities are returned by the P2S model. For each case, a set of synthons are obtained by disconnecting bonds between tagged atoms using RDKit[38]. Subsequently, SMILES strings for these synthons are preprocessed (explained in the Methodology) before feeding them as input to the S2R model to predict possible reactants containing these synthons as substructures.

Challenges for translation-based retrosynthetic model

In recent years, the sequence-to-sequence (seq2seq) based generative models have been widely used in the prediction of single-step retrosynthesis because of its low requirements on data processing (for example, not requiring curation of atom-mapped reactions templates) and strong generalization ability. However, it is not an entirely error-free approach. Lu et al. have summarized three types of predictions based on the retro-synthesis model based on translation model:[20]

1. The SMILES for predicted reactants are grammatically invalid. This problem can be resolved with a simple filter. A small amount of invalid SMILES outputs does not pose a severe challenge.
2. The SMILES for predicted reactants are grammatically valid and chemically plausible, yet the predicted reactants are not identical to the ground-truth reactants specified in the data set. As we know, each molecule may contain multiple reaction centers yielding distinct set of reactants.
3. The SMILES for predicted reactants are valid, but the product-reactants pair does not constitute a chemically plausible reaction. This type of error could be reliably identified by checking subgraph isomorphism between the predicted reactants and the input product. Unfortunately, the subgraph isomorphism search consumes a lot of computational resources.

The second type of errors should not be viewed as a real mistake in the context of synthesis planning. Rather, this notion of "error" does expose the fact that there are always multiple valid approaches to synthesize a molecule. Hence, one expects a single-step retrosynthetic model to enumerate as many valid options as possible. Unfortunately, upon close inspection, the seq2seq translation-based models do not exhibit much diversity in their outputs. Furthermore, they also tend to produce the third type of errors as shown in **Figure 2** and **Figure 3**. **Figure 2** presents S-Transformer's Top-6 recommendations of possible reactants for a selected molecule, while an alternative view of these 6 recommendations (in terms of SMILES which directly output by S-Transformer) is presented in **Figure 3**. As clearly shown in these two figures, S-Transformer's predictions is lack of diversity. Furthermore, many predicted reactants are completely unreasonable in the chemical reaction sense. In order to tackle these challenges, we propose the "Mix and Match" and Label and Align" strategies in RetroPrime to alleviate the problems of poor diversity and low chemical validity, respectively. "Mix and Match" strategy is to explicitly take into account of different choices of decompositions and associated sets of synthons. "Label and Align" uses marked tokens to distinguish and align reaction center and conservative groups between synthons and reactants. See the **Methodology** section for further details.

Baseline

We benchmark our method against six baselines, including four template-free and two template-based methods. Specifically, Seq2Seq[20] is a template-free approach that trains an LSTM model to translate the SMILES of target molecules to SMILES of reactants. RetroSim is a template-based method that recommends templates for target molecules based on the molecular similarity between present molecule and the ones in the dataset. S-Transformer is similar to the Seq2Seq translation model but using single-stage transformer instead of LSTM architecture at core. G2Gs[21] and GraphRetro[23] are template-free approach using the graph neural networks to predict retrosynthesis. GLN[4] is a template-based method, which samples templates and reactants jointly form a distribution learned by a conditional graphical model. Since GLN is possibly the most competitive baseline, we mainly draw comparison to it in the following discussions. However, full comparisons against all baselines are also provided in the tables.

Top-N accuracy

We evaluated the method in two datasets, USPTO-50K and USPTO-full, which contain ~ 50, 000 and ~ 950, 000 reaction data, respectively. See **Methodology** for more details of the datasets. For the USPTO-50K dataset, our results are presented in **Table 1** and **Table 2**, respectively. Our method achieves a Top-1 accuracy of 64.8% and 51.4%, when the reaction type is either known or unknown, respectively. Compared with GLN, the state-of-the-art template-based method, our template-free method is superior to GLN when the reaction type is known, and our method also performs comparably to GLN when the reaction type is unknown.

As shown in **Table 3**, our method gains an upper hand to GLN in terms of Top-N in the large-scale experiments. This outcome implies that our method is more robust to noisy data. please see the **Table S2 and Table S3** for additional details.

Finally, we investigate whether our method provides outputs covering a broad range of chemical reactions. This is crucial if these single-step predictors were to be integrated into a multi-step retrosynthetic route planning. As the setting of unknown-reaction-type is more natural for this purpose, we choose this setting and compare our method against the S-Transformer (as both approaches mainly use Transformer to make predictions). This diversity estimation, based on the second metrics introduced in **Section Reaction Diversity**, is shown in **Table 4**. In this case, it is straightforward to attribute the enhanced diversity to the decision of further processing all valid decompositions within the Top-3 answers found by the P2S model in the workflow summarized in **Figure 1**. In addition, we visualize some typical predicted outcomes given by RetroPrime and S-Transformer. As show in **Figure 4**, RetroPrime generates more diverse results, comparing to the baseline models. This diversity comes from the 'Mix and Match' strategy described in **Section Mix and Match**. Additional results are provided in **Supplementary Information section S2**.

The effects of the "Label and Align" strategy

Recall that we did two things while building the S2R dataset. We align input-output sequences and mark atoms with extra labels. In this section, we attempt to elucidate benefits these efforts provide.

we designed experiments to clarify the benefits of these efforts. In this experiment, we train a modified Transformer that is asked to translate synthons to targets in canonical SMILES, i.e. without sequence alignments and labels. **Table 5** and **6** compares the outcome of the original experiment (as depicted in **Figure 1**) and the new experiment with the S2R model replaced with this newly trained one. The results of Top-1 of the original experiment are 4.6% more accurate than the modified one. This accuracy gain for the Top-1 result is 3.0% when the reaction type is unknown. Moreover, the accuracy gap widens between the two experiments when the comparison is expanded to consider Top-10 results, which is 5.7% and 3.6%, respectively, when the reaction is either known or unknown.

In addition to increasing Top-N accuracy, we further elaborate on more subtle effects brought upon by the labels. It is easy to corroborate that not all outputs of grammatically valid SMILES by a Transformer model are chemically plausible, i.e. the input-output pair does not constitute a valid chemical reaction.

To estimate how many chemically implausible but grammatically valid SMILES are outputted by RetroPrime, we propose to use a forward reaction predictor to diagonalize potential errors. This verification method is inspired by Schwaller et al.[35]. In short, we feed the predictions results (i.e. reactants) of our retrosynthetic method to a forward reaction prediction model, the Molecular Transformer[34]. If the forward model predicts correctly the product molecule within Top-5 choices, then the retrosynthesis is deemed successful. Without taking into account of chirality, the USPTO-MIT mixed version of the Molecular Transformer reaches 94.2%[34] for the Top-5 accuracy.

Results of this scrutiny on chemical validity of our method are summarized in **Table 7** and **8**. Recall that our test set consists of 5,006 cases. For each retrosynthetic prediction, we use Top-10 choices for the forward-reaction test. Based on these results, our method yields slightly more grammatically invalid SMILES in comparison to the modified experiment in which the S2R model is trained with input-output pair given in canonical SMILES without extra labels. However, the potential number of chemically implausible cases are significantly reduced for our proposed method regardless of whether the reaction class is given as part of the input. Since filtering out chemically implausible yet grammatically valid SMILES is significantly trickier, it is certainly suggestive that our method is superior to the modified experiment.

Clearly, our two-stage method has significantly ameliorated this deficiency of the rudimentary workflow using a single Transformer in an end-to-end fashion that directly translates a product molecule into a batch of reactants.

Generalizability across chemical space

We conduct a simple experiment to investigate whether RetroPrime (using a chemist's two-stage strategy) can generalize better than a standard end-to-end machine-learning approach across chemical space. We selected RetroSim and S-Transformer as baselines for this comparison. Using the training set of USPTO-50K (40004 reaction records) as the chemical knowledge base, we tested the Top-n accuracy of these three models on a test set comprising 50,000 reaction records, randomly drawn from USPTO-full minus the USPTO-50K training set. The results are shown in **Table 9**. In principle, USPTO-full contains a lot more molecules that are not similar to the ones in USPTO-50K. The results in **Table 9** show that RetroPrime exhibits better generalizability. While S-Transformer, being also a template-free method, performs better than RetroSim, the advantage seems to diminish as the number of predictions increases. This is, however, not the case with RetroPrime.

Conclusion

In summary, we propose a new Transformer-based method, RetroPrime, to tackle retrosynthesis. RetroPrime not only delivers a comparable performance (in terms of Top-N accuracy) to all state-of-the-art and data-driven methods with the standard USPTO-50K dataset, but it outperforms the best template-based method GLN for the large dataset USPTO-full, comprising million reaction records, by a non-trivial margin of 4.8%. Note that this is one of the only two assessments on a Transformer-based model with a large-scale dataset. The experiment in **Section Generalizability across chemical space** further highlight RetroPrime's generalizability across chemical space. These encouraging results seems to concur with an earlier observation[39] that Transformer-based predictions possess excellent generalizability and robustness.

However, it is easy to show that Transformer suffers from two severe deficiencies: (1) lack of reaction diversity and (2) hard-to-detect chemically implausible solutions. Without further improvements on these two issues, one cannot trust Transformer's outputs beyond the first few ones. In this work, we make conscious efforts to address these challenges by proposing the "mix and match" and the "align and label" strategies as part of RetroPrime two-stage workflow, inspired by a chemist's approach to retrosynthesis. While improvements are substantial as reported, further innovations are urgently desired.

Given vast amount of chemical reaction data and new knowledges are generated on a daily basis, the benefits of building a reliable template-free method are obvious. Hopefully, without having to be explicitly trained on all reaction templates, these modern machine-learning methods can generalize more easily and guide us toward better synthetic routes.

Methodology

Data Preparation

To train the two transformers in **Figure 1**, we generate two new datasets by processing information derived from the publicly available reaction dataset USPTO-50K, which contains ~ 50,000 records of atom-mapped reactions that have been classified into ten distinct reaction types[36]. Following other prior studies, we consider two settings for the predictive task depending on whether the reaction type for each data record is provided as part of the input to the model. Furthermore, we adopt the same training/validation/test split as reported in Coley et al[17], which recommends a split of 80%/10%/10% of 50K reactions. **Table 10** succinctly summarizes the USPTO-50K dataset. In these new datasets, each data entry is prepared in the format of <source>-<output> pair, following the standard data format for NLP tasks. Further details on these datasets are elaborated thoroughly in the following sections.

Reaction-Center dataset generation

For each atom-mapped reaction record in the USPTO-50K, we analyze and label the essential atoms of the product molecule involved in a reaction. The P2S model is trained to identify these tagged atoms for each reaction. Hence, the source of the reaction-center dataset is the canonical SMILES of the product, and the target is the same canonical SMILES with tags added to the reactive atoms.

To prepare this dataset, we consider 4 distinct tags, each implies a very specific instruction set to generate synthons. We utilize the *molAtomMapNumber* attribute in RDKit to help with the tagging. The definitions for these 4 tags are summarized below, and further details these 4 tags are summarized below, and further details may be found in **Figure 5** and **Table 11**.

- Case 1, Tag two atoms. Disconnect bonds between these atoms to form two reactants.
- Case 2, Tag at least two atoms but do not disconnect any bonds. The product itself is a synthon.
- Case 3, Tag one atom. While the product is the only synthon, there must be a leaving group.
- Case 4, Tag multiple atoms. Disconnect bonds between these atoms to form two reactants. Ring-forming reactions fall under this scenario.

Data augmentation for Reaction-Center dataset

There always exists multiple valid SMILES to represent one molecule. It has been reported that NLP models, such as various RNN architectures, tend to perform better for applications in the molecular science when the dataset is augmented with same molecules represented in multiple SMILES. In this case, we augment the Reaction-Center dataset by using SMILES enumerator[40] to randomly generate 9 additional SMILES for each canonical one. An illustration is given in **Figure 6**. Note that the source and the target of each data entry only differs by the tags attached to the reactive atoms on the target side; otherwise, the SMILES are exactly the same on every line. **Table 12** provides further details on this dataset.

Synthons-to-Reactants dataset generation

According to the pipeline depicted in **Figure 1**, synthons are generated from the product molecules by following the instructions implied by the tags introduced in **Reaction-Center dataset generation**. These synthons need to be further processed with labels before feeding to the S2R model. The labelling principle is that the reactive atoms (the ones tagged in **Reaction-Center dataset generation**) are marked as 1, the adjacent atoms (connected via chemical bonds) are marked as 2, and the remaining atoms are marked as 3. The labels can be easily added to RDKit's molecule objects by utilizing the *molAtomMapNumber* attribute in RDKit, and the properly ‘labelled’ SMILES can be produced with the RDKit *MolToSmiles* function. This is how we prepare the source (input) part of this dataset. As for the corresponding target (output) part, we take the reactants from the original USPTO-50K dataset and furnish the SMILES with labels according to the above principle. Additionally, the atoms of leaving groups are also marked as 1 for the reactants. Finally, for each synthon-reactant pair, we calculate the edit distance and attempt to minimize it by manipulating the target sequence in order to align the two SMILES strings as closely as possible. As show in **Figure 7**, after alignment, a typical input-output pair in the S2R dataset share a relatively large and identical subsequence. we called this strategy “Label and Align”**.**

Data augmentation for Synthons-to-Reactants dataset

As shown in **Figure 8**, when a SMILES contains multiple entities, we permute the SMILES to generate additional data. For each augmented data entry, we still have to align the source and target sequences to minimize the edit distance. Details of the Synthons-to-Reactants dataset are given in **Table 12**.

Large-scale experiments on USPTO-full

To more comprehensively test our method, we build whole new datasets using the entire set of USPTO-full (1976-Sep2016)[41]. There are 1,808,937 raw records. For reactions involving multiple products, we duplicate the same entry as many times as the number of products. In each copy, we remove all products but one to create additional data with a unique product molecule for the same reaction. After proper data cleaning, we retain a slightly reduced dataset comprising 950K reaction records. We again randomly divide this into a proportion of 80%/10%/10% for training/validation/test set, respectively.

Repeating the procedures given in USPTO-50K dataset processing, we further produce the Reaction-Center dataset and the Synthons-to-Reactants dataset from the USPTO-full. Note, this dataset generation procedure includes the data augmentations described in the previous section. Further details of these large-scale datasets are summarized in **Table 10, 11, 12.**

Models training

As mentioned earlier, we utilize two transformer[33] models in the proposed workflow. Both are built with Open NMT[42]. In this work, we use the following regular expression to separate SMILES into tokens:

Regex = ”(\[[^\]] +|Br?|Cl?|N|O|S|P|F|I|b|c|n|o|s|p|\(|\)|\.|=|#|-|\+|\\\\|\/|:|~|@|\?|>|*|\$|\%[0-9]{2}|[0-9])”

In this scheme, we count the marked tagged atoms as unique tokens. Details of the transformer architecture are provided in **Figure 9**. See **Table S1** for further details such as the hyperparameters and training details in the **Supplementary Information**.

Evaluation metrics

The evaluation metrics we used are slightly different for the two tasks. For the P2S Transformer, it is expected to tag reactive atoms with an appropriate reaction type in a product molecule. Hence, the evaluation metric is a Top-N accuracy with respect to the tagging in the ground truth. For the second Transformer S2R, it is expected to translate synthons to reactants. To boost accuracy, we propose to mark atoms in order to facilitate the alignment of the source and target sequences for this translation task. Hence, one should remove these labels and convert the target sequence (given by the S2R model) back to canonical SMILES before comparing to the ground truth for a given reaction in the USPTO-50K/full dataset.

Reaction Diversity

For reactions with unknown reaction type, we check whether RetroPrime can offer diverse reaction outcomes. We use a reaction type predictor[29] based on typical message-passing graph convolution network[43] to predict the reaction type of a predicted reaction. Take RetroPrime's Top-10 predictions for each test case, we use reaction type predictor to estimate the number of distinct reaction types. Finally, use the average for all test cases as diversity evaluation criterion.

Mix and Match

The P2S model predicts how a molecule can be decomposed into simpler constituents. Various decompositions imply different chemical reactions. In other similar studies, one would simply take synthons for the Top-1 decomposition to make further predictions of reactants. However, we reckon that processing multiple decompositions down the pipeline of **Figure 1** is a simple yet highly effective method to enormously enhance the overall output diversity. We present a schematic to illustrate the "mix-and-match" strategy in **Figure 10**. See **Supplementary Information Figure S1** for further details on the "mix-and-match".

Label and Align

While preparing the S2R dataset, we meticulously minimized the edit distance for the input-output sequences and insert extra labels as detailed in **Synthons-to-Reactants dataset generation.** These efforts aim to expose as much similarity between the source and target sequences as possible and facilitate the learning for the translational model to capture the chemistry behind the data. Indeed, the "Label and Align" strategy not only improves the transformer's overall accuracy but also increase the number of valid outputs, e.g. less appearances of invalid SMILES in output.

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 21775060).

References

1 E. J. Corey, *Angew. Chemie Int. Ed. English*, 1991, **30**, 455–465.

2 M. A. Ott and J. H. Noordik, *Recl. des Trav. Chim. des Pays-Bas*, 1992, **111**, 239–246.

3 M. H. Todd, *Chem. Soc. Rev.*, 2005, **34**, 247–266.

4 H. Dai, C. Li, C. W. Coley, B. Dai and L. Song, 2020, 1–15.

5 A. Cook, A. P. Johnson, J. Law, M. Mirzazadeh, O. Ravitz and A. Simon, *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 2012, **2**, 79–107.

6 W. A. Warr, *Mol. Inform.*, 2014, **33**, 469–476.

7 T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, X. Hou, J. W. Johannes, C. Kreatsoulas, B. Lahue, M. Mathea, G. Mogk, C. A. Nicolaou, A. D. Palmer, D. J. Price, R. I. Robinson, S. Salentin, L. Xing, T. Jaakkola, W. H. Green, R. Barzilay, C. W. Coley and K. F. Jensen, *J. Med. Chem.*, 2020, **63**, 8667–8682.

8 E. J. Corey, *Pure Appl. Chem.*, 1967, **14**, 19–38.

9 W. D. Ihlenfeldt and J. Gasteiger, *Angew. Chemie (International Ed. English)*, 1996, **34**, 2613–2633.

10 O. Engkvist, P. O. Norrby, N. Selmi, Y. hong Lam, Z. Peng, E. C. Sherer, W. Amberg, T. Erhard and L. A. Smyth, *Drug Discov. Today*, 2018, **23**, 1203–1218.

11 F. Feng, L. Lai and J. Pei, *Front. Chem.*, 2018, **6**, 199.

12 S. V. Ley, D. E. Fitzpatrick, R. J. Ingham and R. M. Myers, *Angew. Chemie - Int. Ed.*, 2015, **54**, 3449–3464.

13 D. Caramelli, J. Granda, D. Cambié, H. Mehr, A. Henson and L. Cronin, , DOI:10.26434/chemrxiv.12924968.v1.

14 F. Häse, L. M. Roch and A. Aspuru-Guzik, *Trends Chem.*, 2019, **1**, 282–291.

15 C. W. Coley, W. H. Green and K. F. Jensen, *Acc. Chem. Res.*, 2018, **51**, 1281–1289.

16 V. H. Nair, P. Schwaller and T. Laino, *Chimia (Aarau).*, 2019, **73**, 997–1000.

17 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, *ACS Cent. Sci.*, 2017, **3**, 1237–1245.

18 M. H. S. Segler and M. P. Waller, *Chem. - A Eur. J.*, 2017, **23**, 5966–5971.

19 K. Lin, Y. Xu, J. Pei and L. Lai, *Chem. Sci.*, 2020, **11**, 3355–3364.

20 B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender and V. Pande, *ACS Cent. Sci.*, 2017, **3**, 1103–1113.

21 C. Shi, M. Xu, H. Guo, M. Zhang and J. Tang, *arXiv Prepr. arXiv2003.12725*.

22 J. Nam and J. Kim, *arXiv Prepr. arXiv1612.09529*.

23 V. R. Somnath, C. Bunne, C. W. Coley, A. Krause and R. Barzilay, *arXiv Prepr. arXiv2006.07038*.

24 I. V Tetko, P. Karpov, R. Van Deursen and G. Godin, *Nat. Commun.*, 2020, **11**, 1–11.

25 C. W. Coley, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2019, **59**, 2529–2537.

26 J. Law, Z. Zsoldos, A. Simon, D. Reid, Y. Liu, S. Y. Knew, A. P. Johnson, S. Major, R. A. Wade and H. Y. Ando, *J. Chem. Inf. Model.*, 2009, **49**, 593–602.

27 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.

28 S. Hochreiter and J. Schmidhuber, *Neural Comput.*, 1997, **9**, 1735.

29 B. Chen, T. Shen, T. S. Jaakkola and R. Barzilay, *arXiv Prepr. arXiv1910.09688*.

30 S. Zheng, J. Rao, Z. Zhang, J. Xu and Y. Yang, *J. Chem. Inf. Model.*, 2020, **60**, 47–55.

31 P. Karpov, G. Godin and I. V. Tetko, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, 2019, vol. 11731 LNCS, pp. 817–830.

32 A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-Mcleod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.

33 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, in *Advances in Neural Information Processing Systems*, 2017, vol. 2017-Decem, pp. 5999–6009.

34 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

35 P. Schwaller, R. Petraglia and T. Laino, .

36 N. Schneider, N. Stiefl and G. A. Landrum, *J. Chem. Inf. Model.*, 2016, **56**, 2336–2346.

37 C. W. Coley, L. Rogers, W. H. Green and K. F. Jensen, , DOI:10.1021/acscentsci.7b00355.

38 G. Landrum, 2006.

39 Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod and C. R. Butler, *Chem. Commun.*, 2019, **55**, 12152–12155.

40 E. J. Bjerrum, *arXiv Prepr. arXiv1703.07076*.

41 D. Lowe, *URL https//figshare. com/articles/Chemical_ React.*, , DOI:10.6084/m9.figshare.5104873.v1.

42 G. Klein, Y. Kim, Y. Deng, J. Senellart and A. M. Rush, *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Syst. Demonstr.*, 2017, 67–72.

43 W. Jin, C. W. Coley, R. Barzilay and T. Jaakkola, *Adv. Neural Inf. Process. Syst.*, 2017, **2017-Decem**, 2608–2617.

Footnote

Electronic supplementary information (ESI) available. See DOI:

Table 1. USPTO-50K dataset Top-N exact match accuracy when the reaction type is known.

Methods	Top-N accuracy %			
	1	3	5	10
Liu et al. Seq2Seq[20]	37.4	52.4	57.0	61.7
Coley et al. RetroSim[17]	52.9	73.8	81.2	88.1
S-Transformer	57.3	71.6	75.2	78.0
Shi et al. G2Gs[21]	61.0	81.3	**86.0**	88.7
Dai et al. GLN[4]	64.2	79.1	85.2	**90.0**
RetroPrime	64.8	81.6	85.0	86.9
GraphRetro[23]	**67.8**	**82.7**	85.3	87.0

Table 2. USPTO-50K dataset Top-N exact match accuracy when the reaction type is unknown.

Methods	Top-N accuracy %			
	1	3	5	10
Coley et al. RetroSim	37.3	54.7	63.3	74.1
S-Transformer	43.5	59.2	63.9	68.2
Shi et al. G2Gs	48.9	67.6	72.5	75.5
Dai et al. GLN	52.5	69.0	75.6	83.7
RetroPrime	51.4	70.8	74.0	76.1
GraphRetro	**63.8**	**80.5**	**84.1**	**85.9**

Table 3. USPTO-full dataset Top-N exact match accuracy when the reaction type is unknown.

Methods	Top-N accuracy %			
	1	3	5	10
Coley et al. RetroSim	32.8	-	-	56.1
Dai et al. GLN	39.3	-	-	63.7
RetroPrime	**44.1**	**59.1**	**62.8**	**68.5**

Table 4. Reaction type analysis on USPTO-50K dataset when the reaction type is unknown.

Methods	Reaction type/product
S-Transformer	1.74
RetroPrime	**2.40**

Table 5. Compare the Top-N accuracy of the two methods in the S2R stage when the reaction type is known. Both methods use the same P2S model predicted results.

S2R Methods	Pipeline Top-N accuracy %			
	1	3	5	10
Marked smiles	**64.8**	**81.6**	**85.0**	**86.9**
Canonical smiles	60.2	75.2	78.8	81.2

Table 6. Compare the Top-N accuracy of the two methods in the S2R stage when the reaction type is unknown. Both methods use the same P2S model predicted results.

S2R Methods	Pipeline Top-N accuracy %			
	1	3	5	10
Marked smiles	**51.4**	**70.8**	**74.0**	**76.1**
Canonical smiles	48.4	66.2	70.0	72.5

Table 7. Compare the forward check Top-N accuracy in USPTO-50K test dataset prediction results when the reaction type is known.

Methods	All predictions	Grammatically valid predictions	Forward Check Top-N accuracy %		
			1	3	5
RetroPrime (S2R Marked smiles)	50060	48053	**45.2**	**53.5**	**55.6**
RetroPrime **(**S2R Canonical smiles)	50060	48637	33.7	40.4	42.3
S-Transformer	50060	47121	32.9	39.5	41.4

Table 8. Compare the forward check Top-N accuracy in USPTO-50K test dataset prediction results when the reaction type is unknown.

Methods	All predictions	Grammatically valid predictions	Forward Check Top-N accuracy %		
			1	3	5
RetroPrime (S2R Marked smiles)	50060	49786	**46.8**	**55.9**	**58.3**
RetroPrime **(**S2R Canonical smiles)	50060	49790	42.0	49.7	51.8
S-Transformer	50060	48004	36.4	43.7	45.8

Table 9. Generalization ability test results. The training set of the three methods are the training set data of USPTO-50K, and the test set 50,000 data are randomly selected from USPTO-full.

Methods	Top-N accuracy %			
	1	3	5	10
RetroSim	18.6	27.9	30.8	32.1
S-Transformer	21.9	30.5	33.0	34.6
RetroPrime	**24.4**	**34.8**	**37.2**	**40.7**

Table 10. USPTO-50K/full dataset information.

Dataset	Count	
	USPTO-50K	USPTO-full
Tran	40004	757473
Val	5000	94688
Test	5006	94696
Reaction types	10	None

Table 11. The distributions for the 4 tags in the Reaction-Center prediction dataset.

Tag	Count	
	USPTO-50K	USPTO-full
1	34366	503525
2	2912	78026
3	11606	137480
4	1126	227830

Table 12. USPTO-50K/full Reaction-Center prediction dataset (P2S) and Synthons-to-Reactants dataset (S2R) information.

Setting	Count			
	USPTO-50K		USPTO-full	
	P2S	S2R	P2S	S2R
Train	400040	68373	7574730	1576929
Val	5000	5000	94688	94688
Test	5006	5006	94696	94696
Reaction types	10		None	

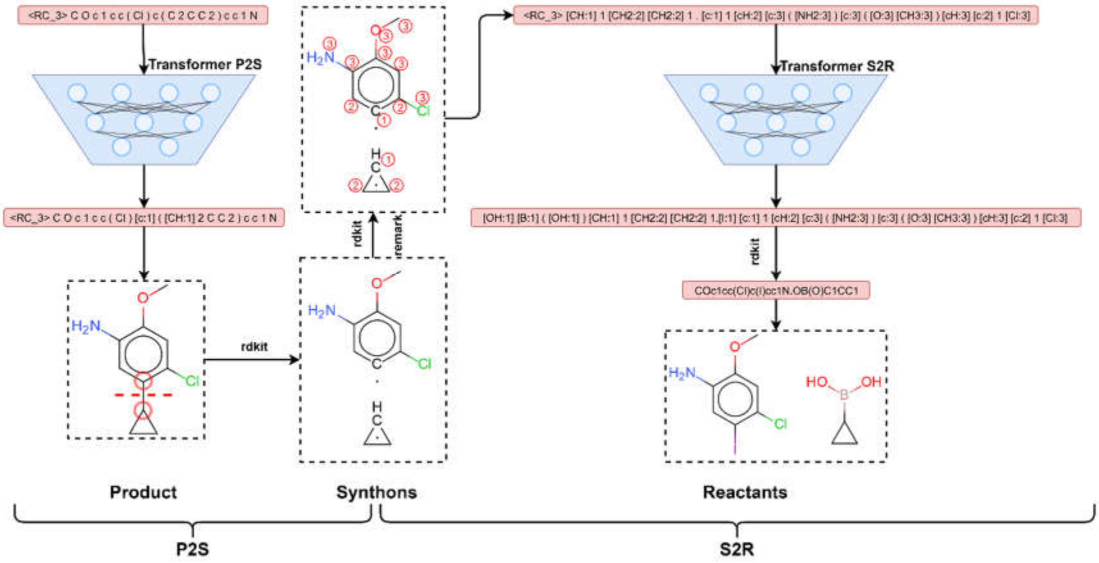
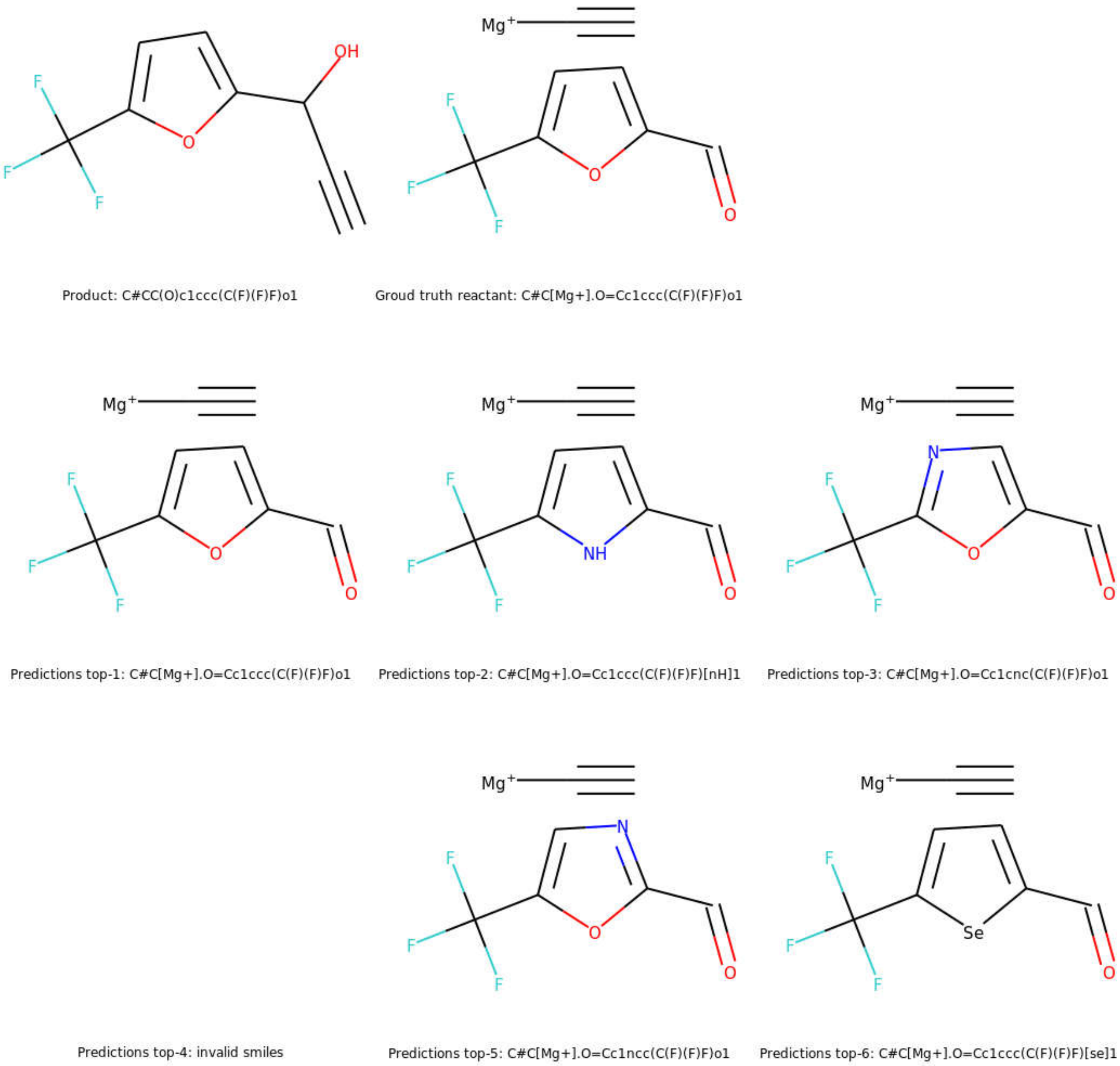

<RC_3> C O c 1 c c (Cl) c (C 2 C C 2) c c 1 N
Transformer P2S
<RC_3> C O c 1 c c (Cl) [c:1] ([CH:1] 2 C C 2) c c 1 N
H2N
O
Cl
rdkit
H2N
O
Cl
C
H
C
rdkit
remark
H2N
O
Cl
C
H
C
<RC_3> [CH:1] 1 [CH2:2] [CH2:2] 1 . [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
Transformer S2R
[OH:1] [B:1] ([OH:1]) [CH:1] 1 [CH2:2] [CH2:2] 1.[I:1] [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
rdkit
COc1cc(Cl)c(I)cc1N.OB(O)C1CC1
H2N
O
Cl
I
HO
B
OH
Product
Synthons
Reactants
P2S
S2R

Figure 1. Method pipeline. First, the canonical SMILES of the product is input into the Transformer P2S to obtain the product SMILES with the reaction center tag. The second step is to use RDKit to disconnect the bond between the tagged atoms (if the tag is a disconnected mark). The third step, remark, and optimized sequence with RDKit, the reaction center is placed at the front of the sequence. The fourth step is to input the synthons into the Transformer S2R to predict the corresponding reactants. Finally, use RDKit to remove the mark and convert it into canonical SMILES.



Product: C#CC(O)c1ccc(C(F)(F)F)o1

Groud truth reactant: C#C[Mg+].O=Cc1ccc(C(F)(F)F)o1

Predictions top-1: C#C[Mg+].O=Cc1ccc(C(F)(F)F)o1

Predictions top-2: C#C[Mg+].O=Cc1ccc(C(F)(F)F)[nH]1

Predictions top-3: C#C[Mg+].O=Cc1cnc(C(F)(F)F)o1

Predictions top-4: invalid smiles

Predictions top-5: C#C[Mg+].O=Cc1ncc(C(F)(F)F)o1

Predictions top-6: C#C[Mg+].O=Cc1ccc(C(F)(F)F)[se]1

Figure 2. Visualization of outputs by the S-Transformer. The first row gives an input molecule and the associated reactants (as specified in the test dataset), and the next two rows give the Top-6 predicted results. In this example, S-Transformer correctly returns the ground-truth reactants in its first attempt, but every subsequent output is unreasonable due to the atomic changes on the conjugated five-member ring in one of the reactants. (The result of Top-4 is grammatically invalid. The SMILES marked in the figure are all canonical SMILES sequences. These sequences directly output by the S-Transformer are shown in Figure 3.)

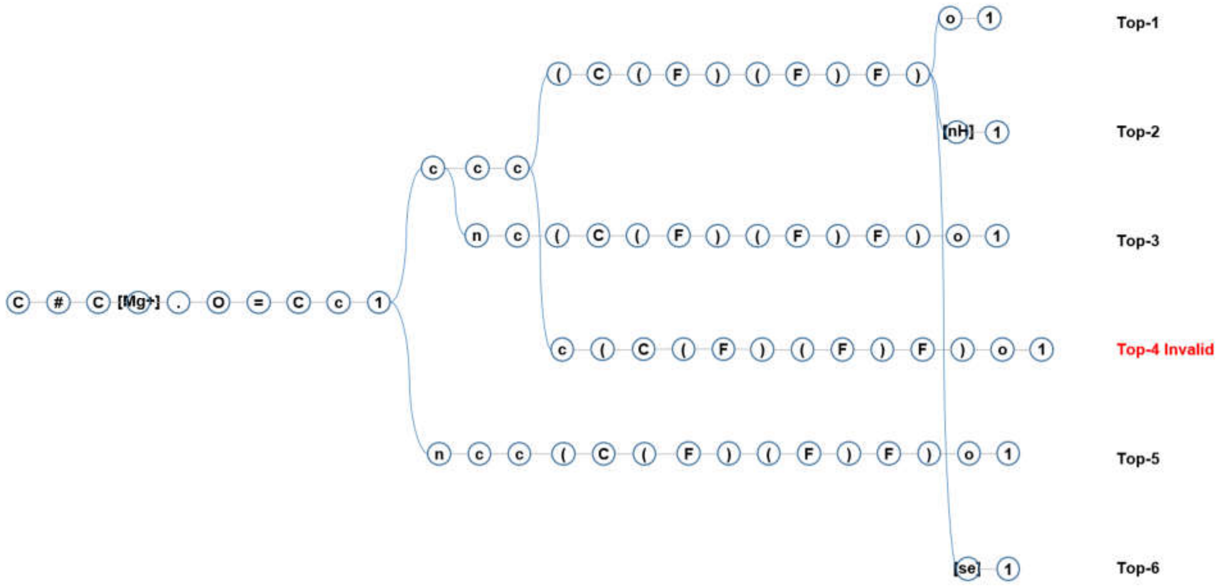

C # C [Mg+] . O = C c 1
c c c (C (F) (F) F) o 1
Top-1
[nH] 1
Top-2
n c (C (F) (F) F) o 1
Top-3
c (C (F) (F) F) o 1
Top-4 Invalid
n c c (C (F) (F) F) o 1
Top-5
[se] 1
Top-6

Figure 3. The sequences directly output by the S-Transformer. This illustration manifests that most sequences look highly similar. Most predictions indicate that the transformer model does not understand reaction center and conservative groups. In most case, when the predicted SMILES are different from the ground truth, the molecules represented by those SMILES and the input products cannot form a reasonable chemical reaction.

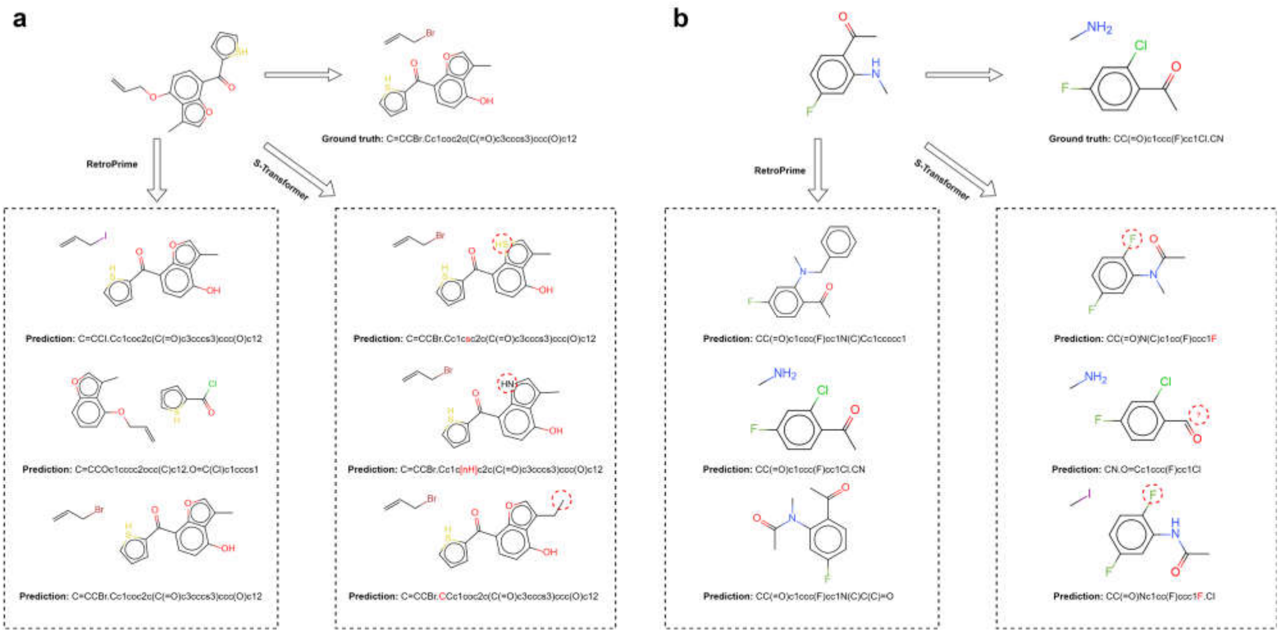

a
Ground truth: C=CCBr.Cc1coc2c(C(=O)c3cccs3)ccc(O)c12
RetroPrime
S-Transformer
Prediction: C=CCI.Cc1coc2c(C(=O)c3cccs3)ccc(O)c12
Prediction: C=CCOc1cccc2occ(C)c12.O=C(Cl)c1cccs1
Prediction: C=CCBr.Cc1coc2c(C(=O)c3cccs3)ccc(O)c12
Prediction: C=CCBr.Cc1csc2c(C(=O)c3cccs3)ccc(O)c12
Prediction: C=CCBr.Cc1c[nH]c2c(C(=O)c3cccs3)ccc(O)c12
Prediction: C=CCBr.CCc1coc2c(C(=O)c3cccs3)ccc(O)c12
b
Ground truth: CC(=O)c1ccc(F)cc1Cl.CN
RetroPrime
S-Transformer
Prediction: CC(=O)c1ccc(F)cc1N(C)Cc1ccccc1
Prediction: CC(=O)c1ccc(F)cc1Cl.CN
Prediction: CC(=O)c1ccc(F)cc1N(C)C(C)=O
Prediction: CC(=O)N(C)c1cc(F)ccc1F
Prediction: CN.O=Cc1ccc(F)cc1Cl
Prediction: CC(=O)Nc1cc(F)ccc1F.Cl

Figure 4. Comparing the predictions of RetroPrime and S-Transformer when the reaction type is unknown, we can see that RetroPrime can predict a variety of different retrosynthesis schemes while the results generated by the S-Transformer are similar. While many results generated by the S-Transformer are very close to ground truth, but the overall reaction is not chemically plausible.

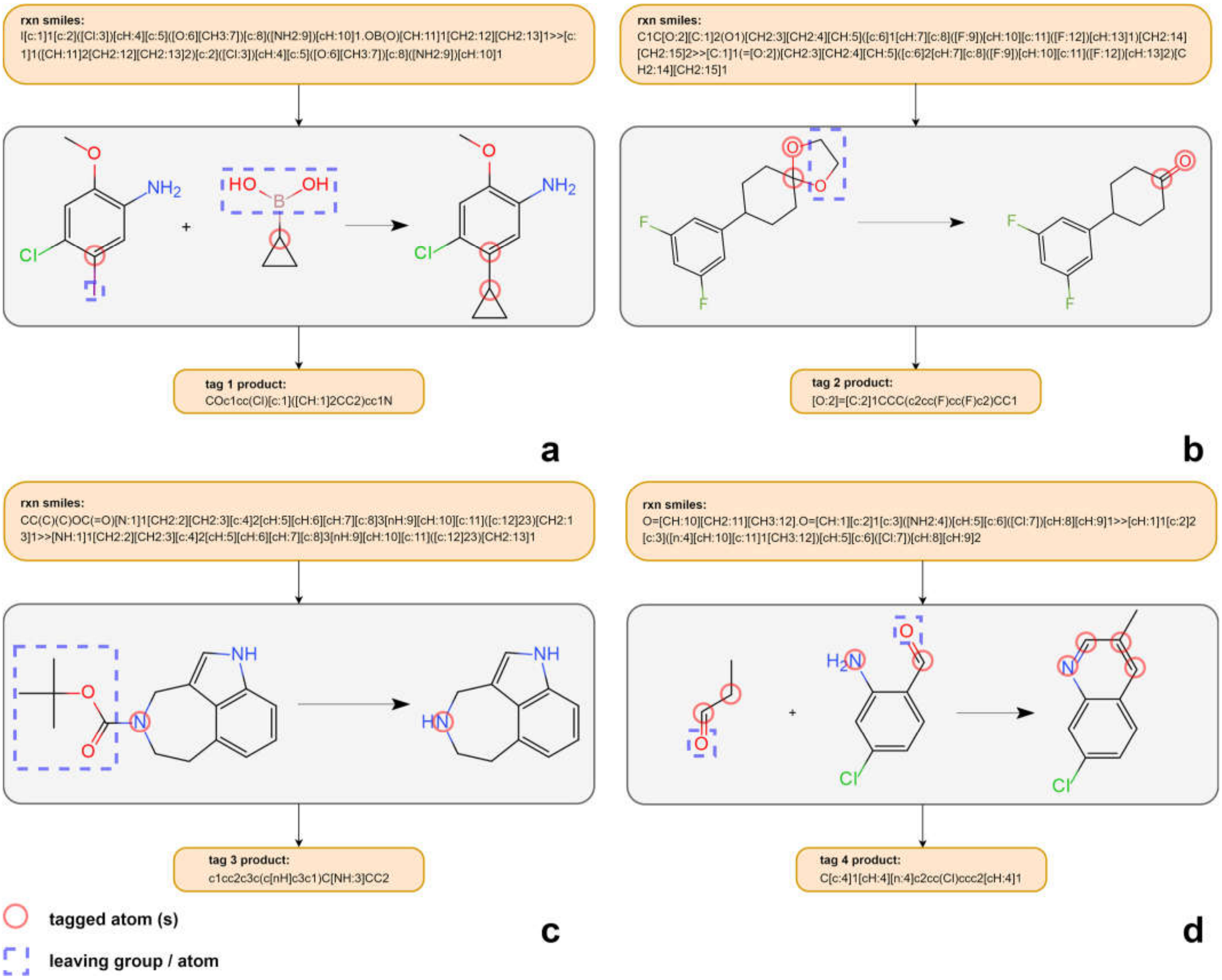
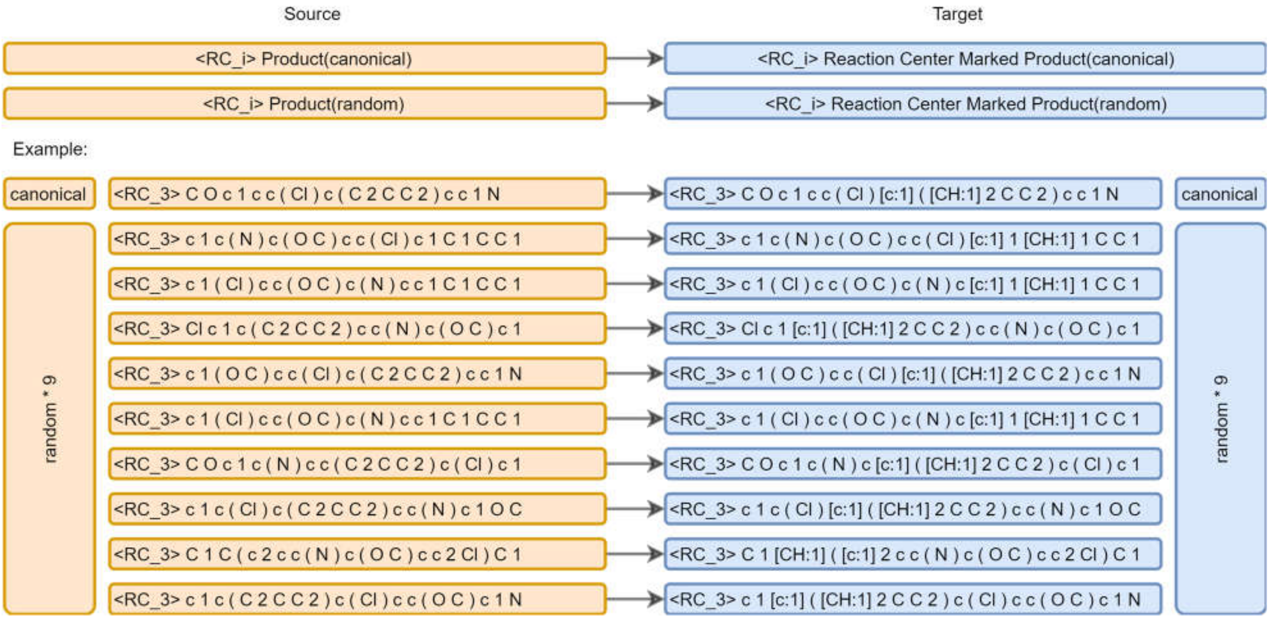

rxn smiles:
I[c:1]1[c:2]([Cl:3])[cH:4][c:5]([O:6][CH3:7])[c:8]([NH2:9])[cH:10]1.OB(O)[CH:11]1[CH2:12][CH2:13]1>>[c:1]1([CH:11]2[CH2:12][CH2:13]2)[c:2]([Cl:3])[cH:4][c:5]([O:6][CH3:7])[c:8]([NH2:9])[cH:10]1
NH2
Cl
I
HO
OH
B
+
tag 1 product:
COc1cc(Cl)[c:1]([CH:1]2CC2)cc1N
a
rxn smiles:
C1C[O:2][C:1]2(O1)[CH2:3][CH2:4][CH:5]([c:6]1[cH:7][c:8]([F:9])[cH:10][c:11]([F:12])[cH:13]1)[CH2:14][CH2:15]2>>[C:1]1(=[O:2])[CH2:3][CH2:4][CH:5]([c:6]2[cH:7][c:8]([F:9])[cH:10][c:11]([F:12])[cH:13]2)[CH2:14][CH2:15]1
O
F
tag 2 product:
[O:2]=[C:2]1CCC(c2cc(F)cc(F)c2)CC1
b
rxn smiles:
CC(C)(C)OC(=O)[N:1]1[CH2:2][CH2:3][c:4]2[cH:5][cH:6][cH:7][c:8]3[nH:9][cH:10][c:11]([c:12]23)[CH2:13]1>>[NH:1]1[CH2:2][CH2:3][c:4]2[cH:5][cH:6][cH:7][c:8]3[nH:9][cH:10][c:11]([c:12]23)[CH2:13]1
NH
O
N
HN
tag 3 product:
c1cc2c3c(c[nH]c3c1)C[NH:3]CC2
c
rxn smiles:
O=[CH:10][CH2:11][CH3:12].O=[CH:1][c:2]1[c:3]([NH2:4])[cH:5][c:6]([Cl:7])[cH:8][cH:9]1>>[cH:1]1[c:2]2[c:3]([n:4][cH:10][c:11]1[CH3:12])[cH:5][c:6]([Cl:7])[cH:8][cH:9]2
O
+
H2N
Cl
N
tag 4 product:
C[c:4]1[cH:4][n:4]c2cc(Cl)ccc2[cH:4]1
d
tagged atom (s)
leaving group / atom

Figure 5. The interpretation of the reaction center tag. (a) Case 1, Tag two atoms. Disconnect bonds between these atoms to form two reactants, (b) Case 2, Tag at least two atoms but do not disconnect any bonds. The product itself is a synthon. (c) Case 3, Tag one atom. While the product is the only synthon, there must be a leaving group, this tag is a non-disconnected mark, and the given product is a synthon, and (d) Case 4, Tag multiple atoms. Disconnect bonds between these atoms to form two reactants. Ring-forming reactions fall under this scenario.



	Source	Target	
	<RC_i> Product(canonical)	<RC_i> Reaction Center Marked Product(canonical)	
	<RC_i> Product(random)	<RC_i> Reaction Center Marked Product(random)	
Example:			
canonical	<RC_3> C O c 1 c c (Cl) c (C 2 C C 2) c c 1 N	<RC_3> C O c 1 c c (Cl) [c:1] ([CH:1] 2 C C 2) c c 1 N	canonical
random * 9	<RC_3> c 1 c (N) c (O C) c c (Cl) c 1 C 1 C C 1	<RC_3> c 1 c (N) c (O C) c c (Cl) [c:1] 1 [CH:1] 1 C C 1	random * 9
	<RC_3> c 1 (Cl) c c (O C) c (N) c c 1 C 1 C C 1	<RC_3> c 1 (Cl) c c (O C) c (N) c [c:1] 1 [CH:1] 1 C C 1	
	<RC_3> Cl c 1 c (C 2 C C 2) c c (N) c (O C) c 1	<RC_3> Cl c 1 [c:1] ([CH:1] 2 C C 2) c c (N) c (O C) c 1	
	<RC_3> c 1 (O C) c c (Cl) c (C 2 C C 2) c c 1 N	<RC_3> c 1 (O C) c c (Cl) [c:1] ([CH:1] 2 C C 2) c c 1 N	
	<RC_3> c 1 (Cl) c c (O C) c (N) c c 1 C 1 C C 1	<RC_3> c 1 (Cl) c c (O C) c (N) c [c:1] 1 [CH:1] 1 C C 1	
	<RC_3> C O c 1 c (N) c c (C 2 C C 2) c (Cl) c 1	<RC_3> C O c 1 c (N) c [c:1] ([CH:1] 2 C C 2) c (Cl) c 1	
	<RC_3> c 1 c (Cl) c (C 2 C C 2) c c (N) c 1 O C	<RC_3> c 1 c (Cl) [c:1] ([CH:1] 2 C C 2) c c (N) c 1 O C	
	<RC_3> C 1 C (c 2 c c (N) c (O C) c c 2 Cl) C 1	<RC_3> C 1 [CH:1] ([c:1] 2 c c (N) c (O C) c c 2 Cl) C 1	
	<RC_3> c 1 c (C 2 C C 2) c (Cl) c c (O C) c 1 N	<RC_3> c 1 [c:1] ([CH:1] 2 C C 2) c (Cl) c c (O C) c 1 N	

Figure 6. Reaction center prediction data augmentation. we generated an additional nine pieces of augmentation data in the training set. <RC_i> is the reaction type if applicable, and we use the reaction type in both source and target.

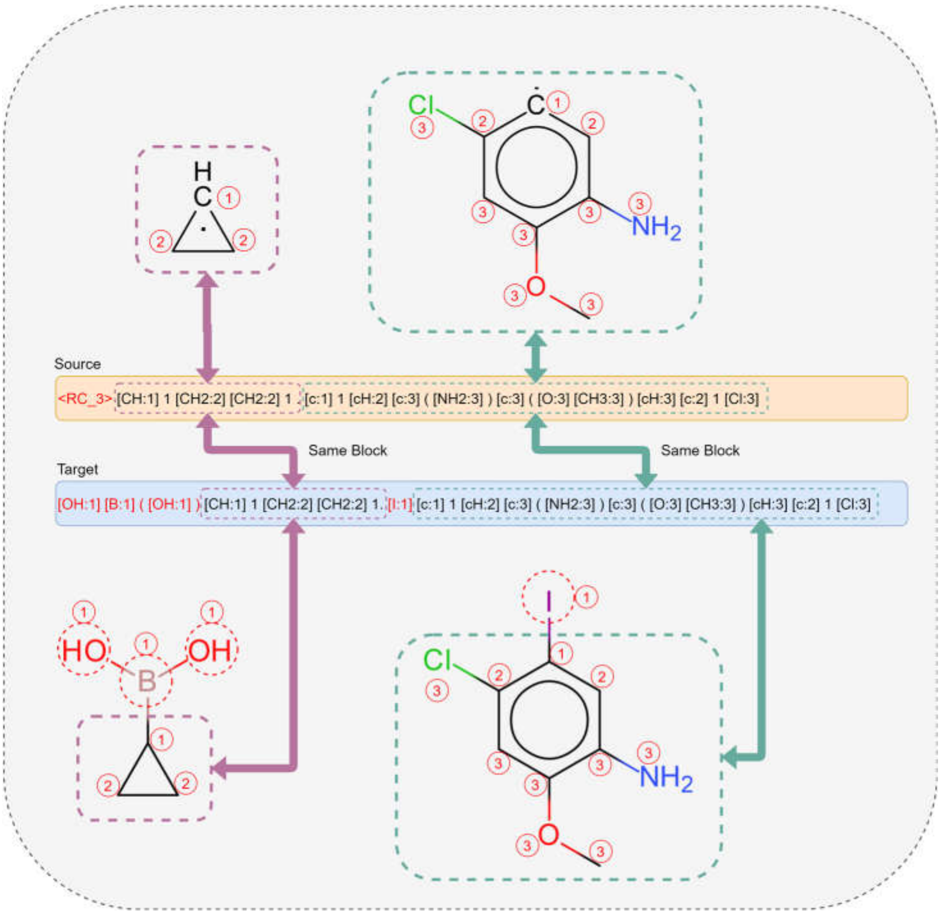

Cl
C
NH2
O
H
C
Source
<RC_3> [CH:1] 1 [CH2:2] [CH2:2] 1 . [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
Same Block
Same Block
Target
[OH:1] [B:1] ([OH:1]) [CH:1] 1 [CH2:2] [CH2:2] 1 . [I:1] [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
HO
B
OH
I
Cl
NH2
O

Figure 7. Label and Align. We use marked SMILES that minimize the editing distance in the S2R stage so that the source and target SMILES have many blocks that are exactly the same.

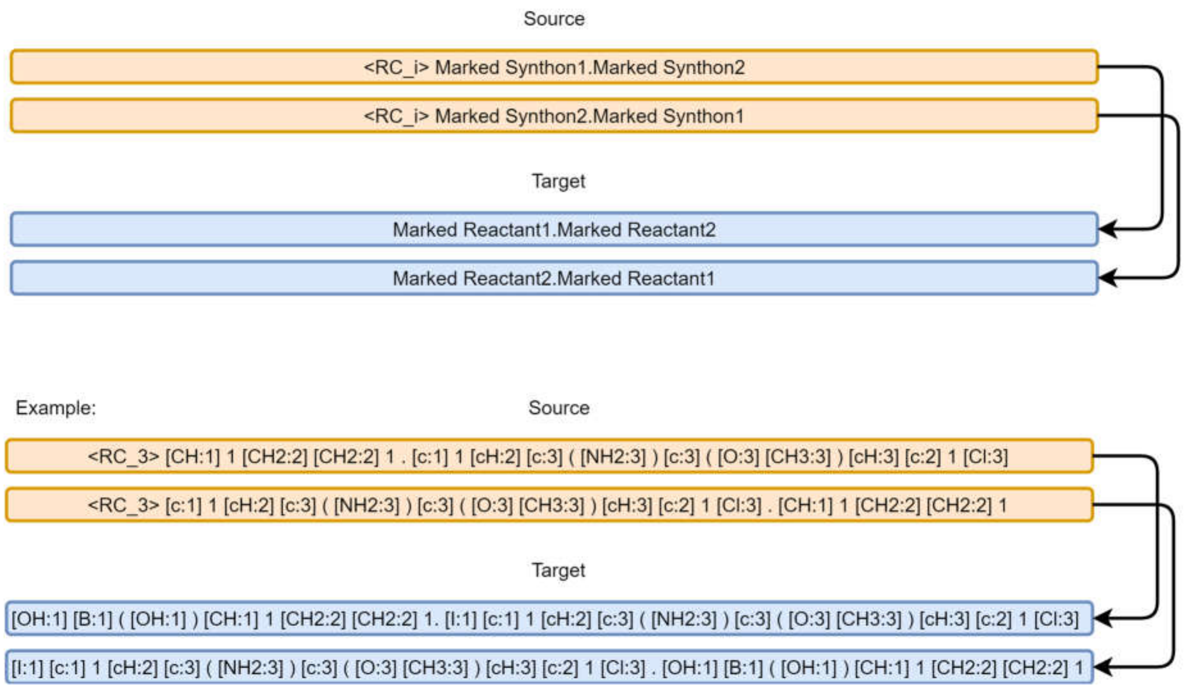

Source
<RC_i> Marked Synthon1.Marked Synthon2
<RC_i> Marked Synthon2.Marked Synthon1
Target
Marked Reactant1.Marked Reactant2
Marked Reactant2.Marked Reactant1
Example:
Source
<RC_3> [CH:1] 1 [CH2:2] [CH2:2] 1 . [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
<RC_3> [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3] . [CH:1] 1 [CH2:2] [CH2:2] 1
Target
[OH:1] [B:1] ([OH:1]) [CH:1] 1 [CH2:2] [CH2:2] 1. [I:1] [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
[I:1] [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3] . [OH:1] [B:1] ([OH:1]) [CH:1] 1 [CH2:2] [CH2:2] 1

Figure 8. Synthons-to-Reactants datasets augmentation. <RC_i> is the reaction type if applicable.

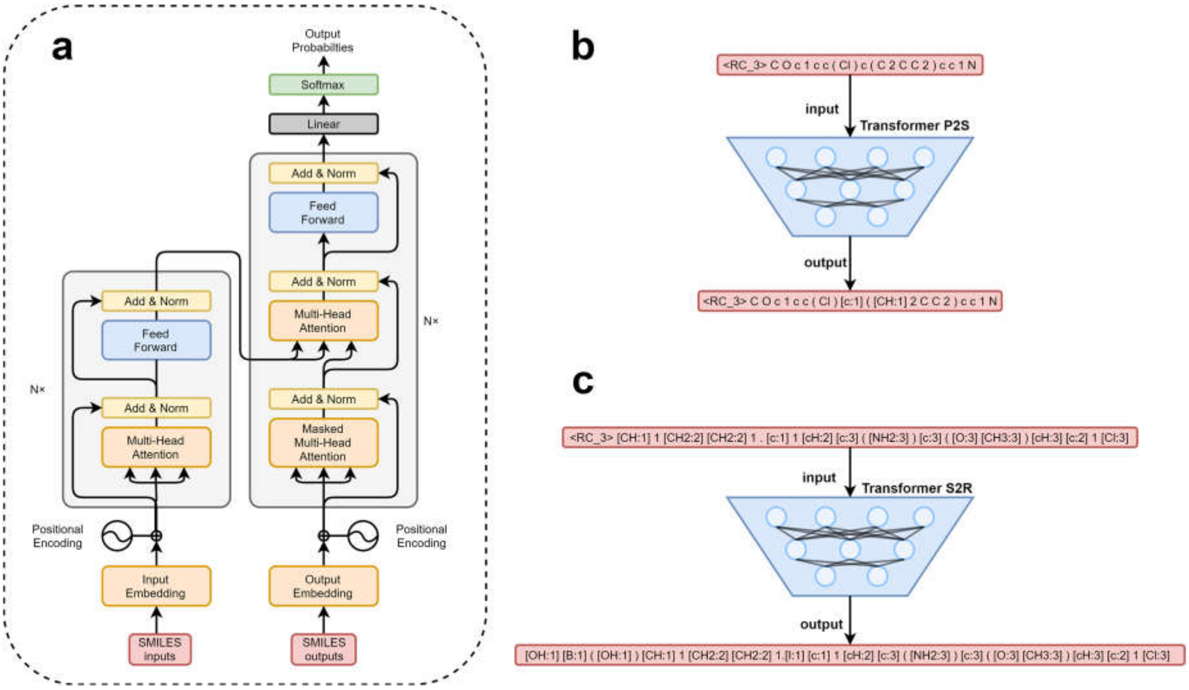

a
Output Probabilties
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
N×
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
N×
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
SMILES inputs
SMILES outputs
b
<RC_3> C O c 1 c c (Cl) c (C 2 C C 2) c c 1 N
input
Transformer P2S
output
<RC_3> C O c 1 c c (Cl) [c:1] ([CH:1] 2 C C 2) c c 1 N
c
<RC_3> [CH:1] 1 [CH2:2] [CH2:2] 1 . [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]
input
Transformer S2R
output
[OH:1] [B:1] ([OH:1]) [CH:1] 1 [CH2:2] [CH2:2] 1.[I:1] [c:1] 1 [cH:2] [c:3] ([NH2:3]) [c:3] ([O:3] [CH3:3]) [cH:3] [c:2] 1 [Cl:3]

Figure 9. Models and their input and output. The parameters of the two models are the same except for the word embedding layer. (a) Transformer P2S and S2R model architecture. (b) Both the input and output of Transformer P2S use canonical SMILES and have reaction type tokens if applicable. (c) Atom symbol and its mark are treated as one token in the input and output of Transformer S2R. Only reaction type token in the input of Transformer S2R if applicable.

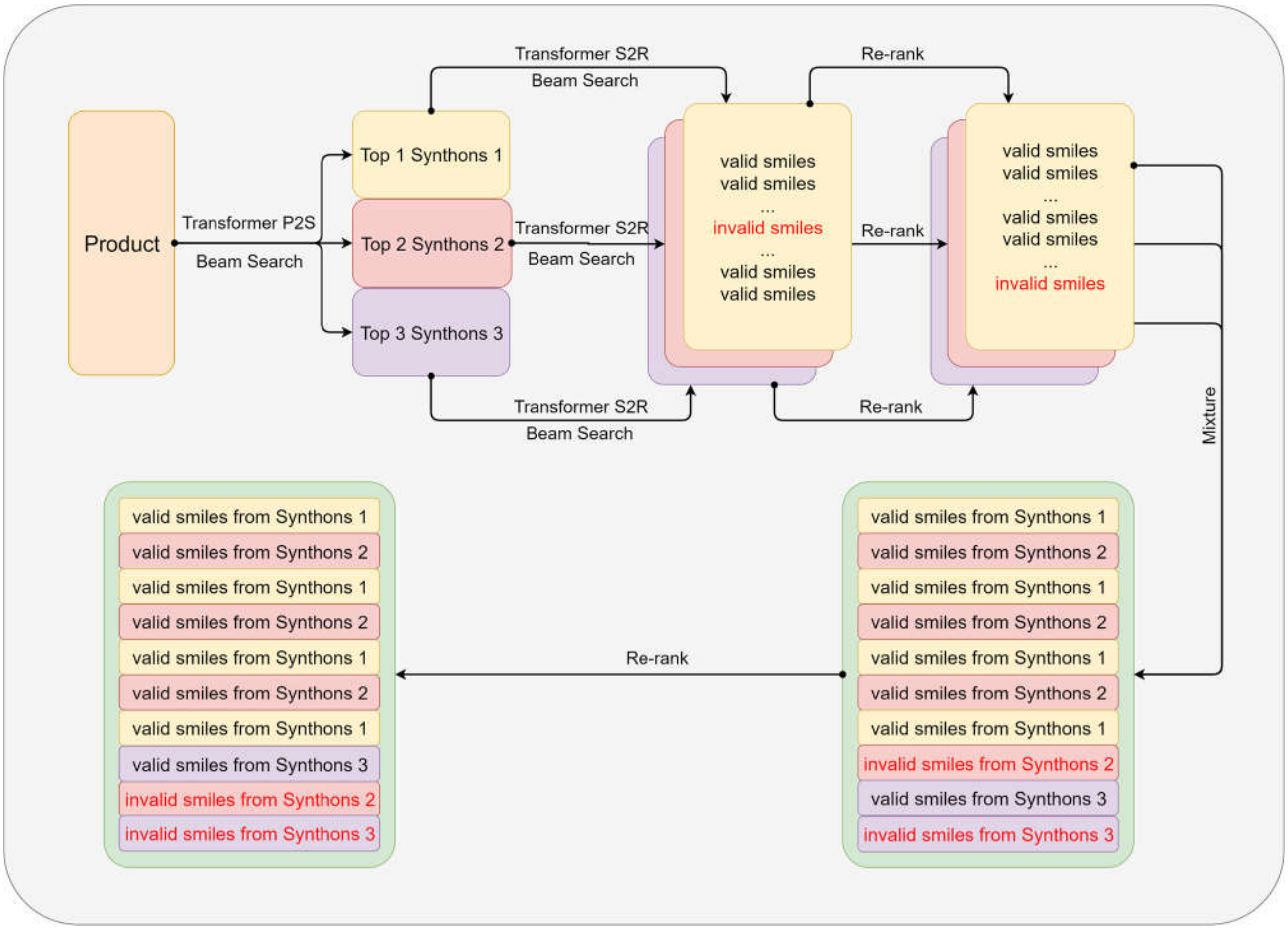
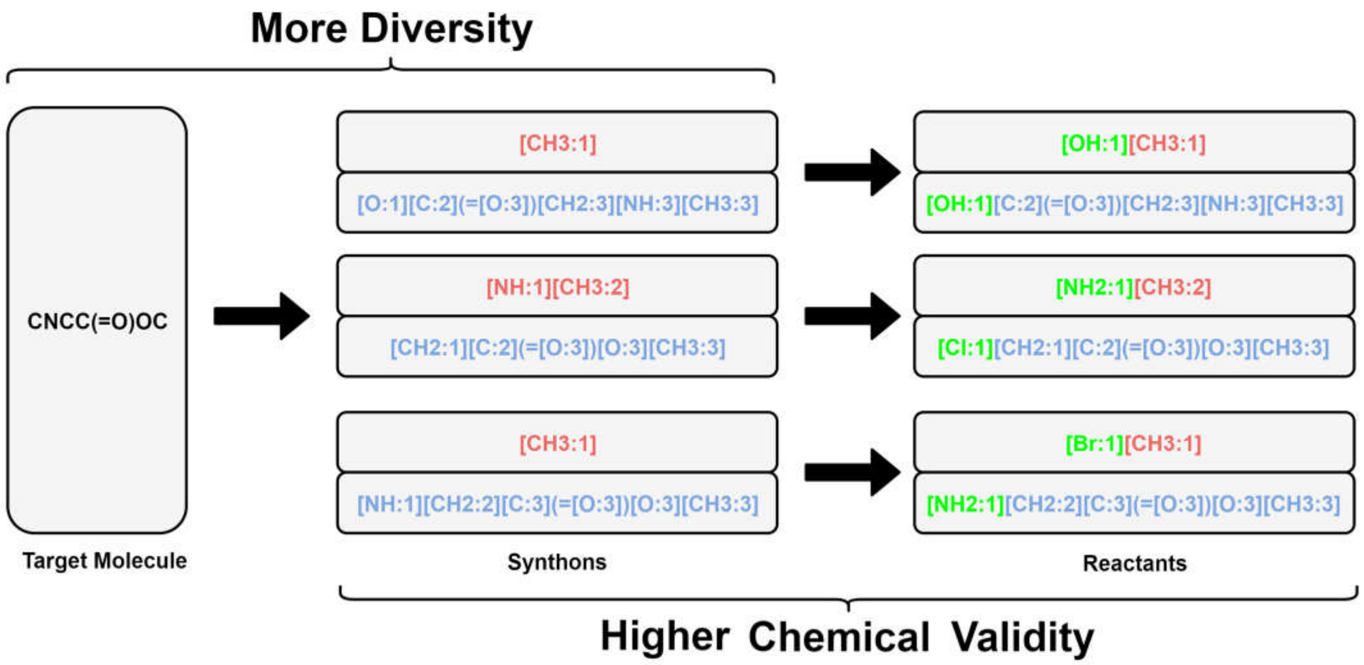

Product
Transformer P2S
Beam Search
Top 1 Synthons 1
Top 2 Synthons 2
Top 3 Synthons 3
Transformer S2R
Beam Search
Transformer S2R
Beam Search
Transformer S2R
Beam Search
valid smiles
valid smiles
...
invalid smiles
...
valid smiles
valid smiles
Re-rank
Re-rank
Re-rank
valid smiles
valid smiles
...
valid smiles
valid smiles
...
invalid smiles
Mixture
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
invalid smiles from Synthons 2
valid smiles from Synthons 3
invalid smiles from Synthons 3
Re-rank
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
valid smiles from Synthons 2
valid smiles from Synthons 1
valid smiles from Synthons 3
invalid smiles from Synthons 2
invalid smiles from Synthons 3

Figure 10. Mix and Match. We select the rank 1-3 synthons predicted by Transformer P2S and send them to Transformer S2R to predict the reactants, and the obtained results are alternately combined. Use the re-rank approach to rank invalid SMILES at the end.


More Diversity
CNCC(=O)OC
Target Molecule
[CH3:1]
[O:1][C:2](=[O:3])[CH2:3][NH:3][CH3:3]
[NH:1][CH3:2]
[CH2:1][C:2](=[O:3])[O:3][CH3:3]
[CH3:1]
[NH:1][CH2:2][C:3](=[O:3])[O:3][CH3:3]
Synthons
[OH:1][CH3:1]
[OH:1][C:2](=[O:3])[CH2:3][NH:3][CH3:3]
[NH2:1][CH3:2]
[Cl:1][CH2:1][C:2](=[O:3])[O:3][CH3:3]
[Br:1][CH3:1]
[NH2:1][CH2:2][C:3](=[O:3])[O:3][CH3:3]
Reactants
Higher Chemical Validity

Graphic abstract