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 RetromPrime can more effectively generalize across chemical space.

1. 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].

At present, purely machine-learning models are classified into two categories[15]: the template-based[4,16,17] and template-free[18,19,20,21,22] methods. A template-based algorithm extracts reaction templates from chemical data[23,24], 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-basd methods can only predict reactions if corresponding templates have been curated in a database[4,16]. With ever growing list of reaction templates, it is certainly desirable to contemplate alternative approaches.

It is often claimed that template-free method 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 Supporting Information, 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.[19] treated the one-step retrosynthesis as a translation task, using SMILES[25] to represent molecules and using an LSTM[26] model, a venerable tool in natural language processing (NLP), to convert SMILES of a product to SMILES of reactant(s). Later on, many

researchers[18,27,28,29,30] adopted more advanced NLP model, the Transformer[31], 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'[32], i.e. predicting a product molecule given a set of reactants and reagents. In another study[30] , Lee et. al. unambiguously demonstrated that generalizability of Transformer model across chemical spaces. Transformer not only performs well in single-step retrosynthesis, but also some researchers have tried to use it in multi-step retrosynthesis method. Lin et al.[18] tried to use the Transformer and Monte-Carlo tree search and found out the retrosynthesis route of four molecules.

While Transformer-based models possess so many desiderata, they suffer from two severe shortcomings: (1) lack of diverse outputs[33] 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.[34] proposed a multifaceted evaluation scheme to replace the Top-N accuracy that could capture these two subtle issues to some extent. 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.[28] 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.[27] 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 section 3.2 and Supporting Information Figure S2.

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[20,22] 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 invokes 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 section 2,

We have not only evaluated our methods on a standard dataset USPTO-50K[35] but also tested on the large-scale USPTO-full[4]. This is one of the few results for template-free methods tested with roughly a million of reaction data records. 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 3.4, 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: Molecular Transformer and RetroSim.

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

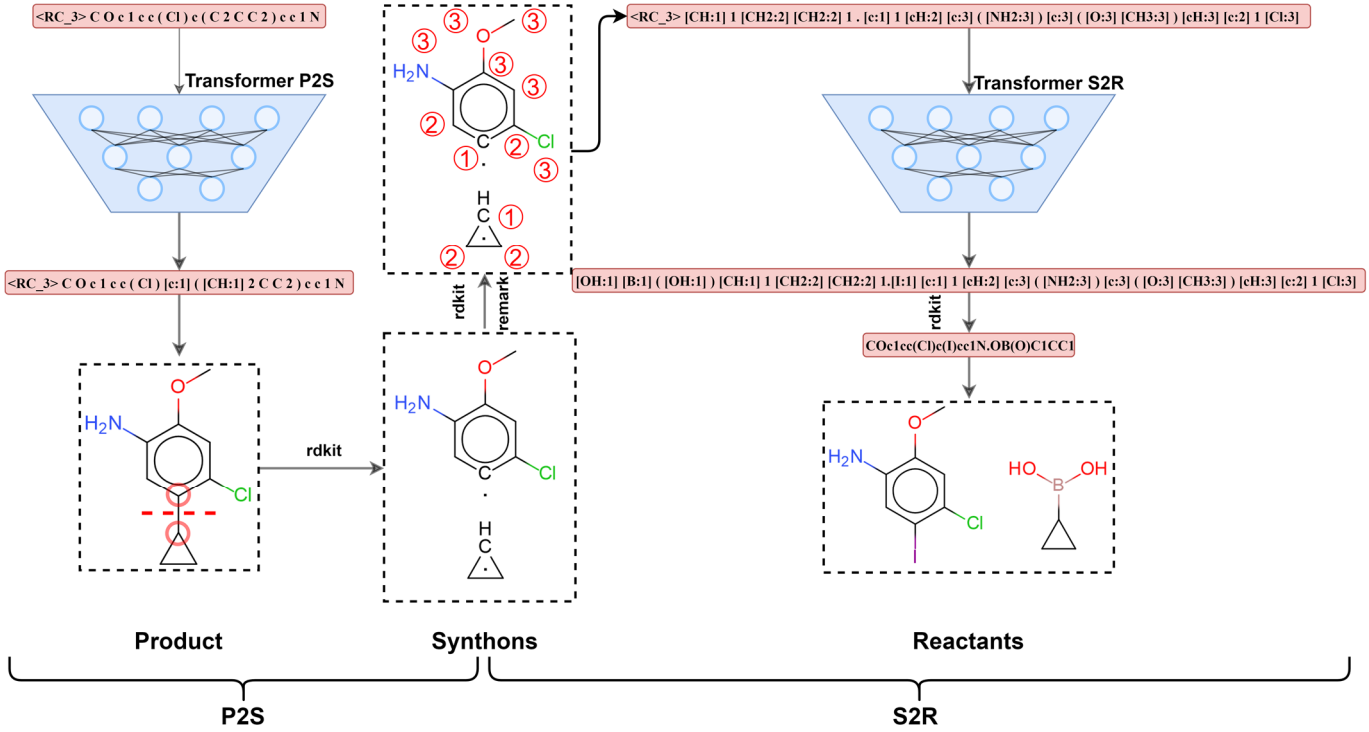

Figure 1. Method pipeline. First, the canonical SMILES of the product is input into the Transformer P2S to obtain the product SMILES with the reaction center tag. The second step is to use RDKit[36] to disconnect the bond between the tagged atoms (if the tag is a disconnected mark). The third step, remark, and optimized sequence with RDKit[36], 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[36] to remove the mark and convert it into canonical SMILES.

methods that, hopefully, may enable fully automated and data-driven retrosynthetic planning of complex molecules in the future.

2. Methodology

2.1. 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 ($P \rightarrow S$). 2. Transform synthons to reactants ($S \rightarrow R$). Both tasks can be accomplished with the help of 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 in the two stages as the product-to-synthons (P2S) model and as 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 be involved in a reaction. Multiple possibilities are returned by the P2S model. For each case, using RDKit, a set of synthons are obtained by disconnecting bonds between tagged atoms. Subsequently, SMILES strings for these synthons are preprocessed (explained below) before feeding them as input to the S2R model to predict possible reactants containing these synthons as substructures.

2.2. 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-

Table 1. 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 2. 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 3. 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	

50K, which contains ~ 50,000 records of atom-mapped reactions that have been classified into ten distinct reaction types[35]. 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[16], which recommends a split of 80%/10%/10% of 50k reactions. Table 1 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.

2.2.1. Reaction-Center dataset

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[36] to help with the tagging. The definitions for

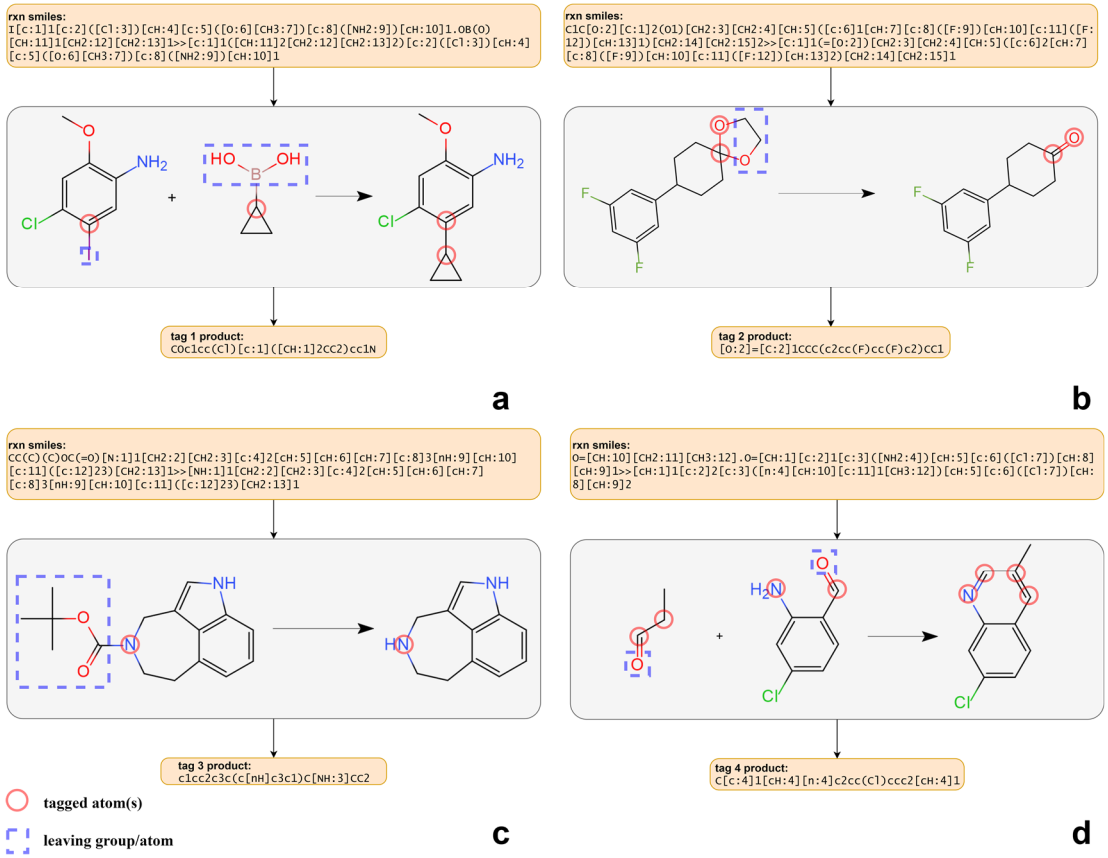

Figure 2. The interpretation of the reaction center tag. (a) Case 1, Tag two atoms. Disconnect bonds between these atoms to form two reactants, (b) Case 2, Tag at least two atoms but do not disconnect any bonds. The product itself is a synthon. (c) Case 3, Tag one atom. While the product is the only 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.

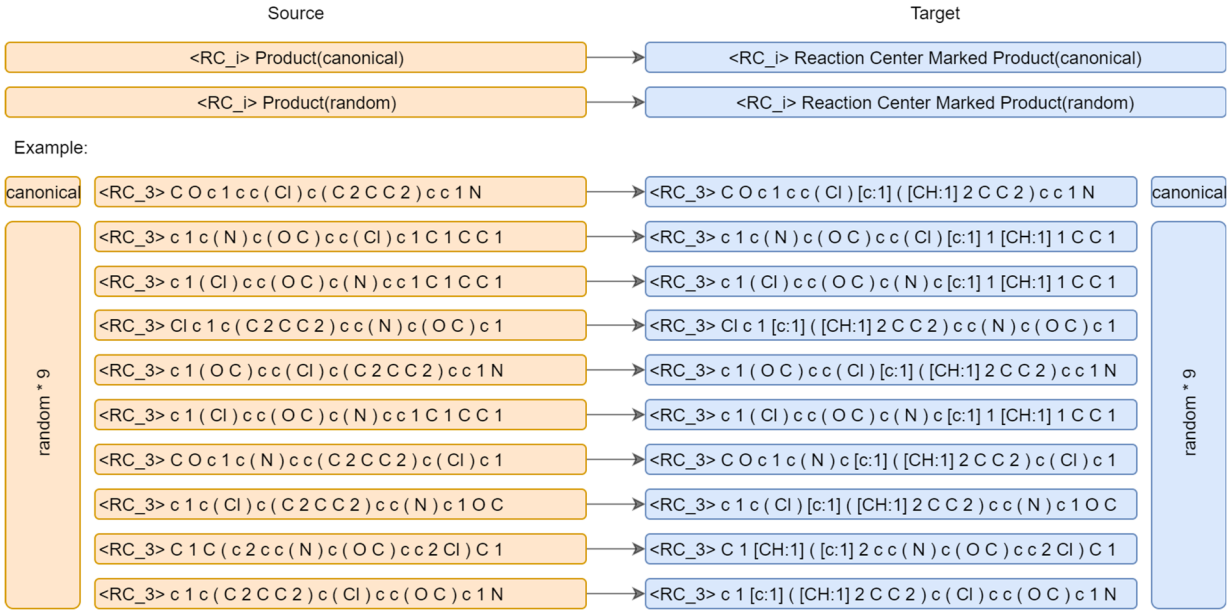

Figure 3. Reaction center prediction data augmentation. we generated an additional nine pieces of augmentation data in the training set. <RC_i> is the reaction type if applicable, and we use the reaction type in both source and target.

these 4 tags are summarized below, and further details may be found in Figure 2 and Table 2.

- 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.

2.2.2. 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[37] to randomly generate 9 additional SMILES for each canonical one. An illustration is given in Figure 3. 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 3 provides further details on this dataset.

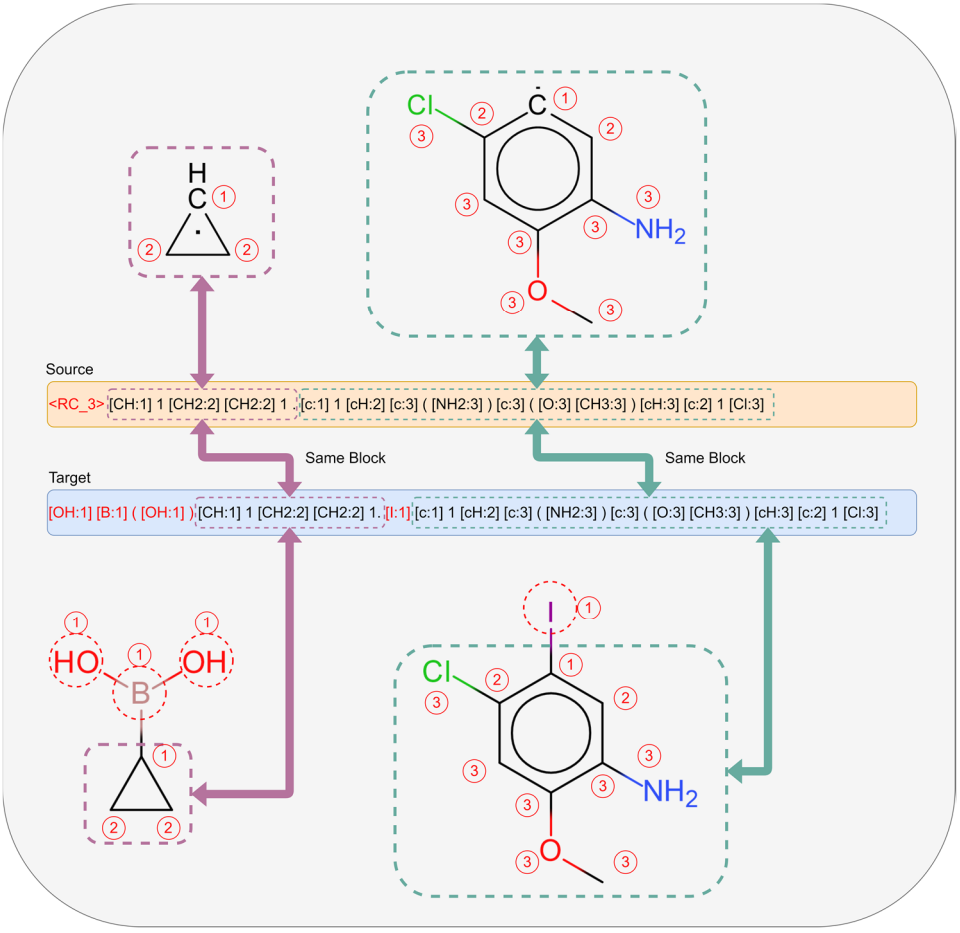

Figure 4. Label and Align. We use marked SMILES that minimize the editing distance in the S2R stage so that the source and target SMILES have many blocks that are exactly the same.

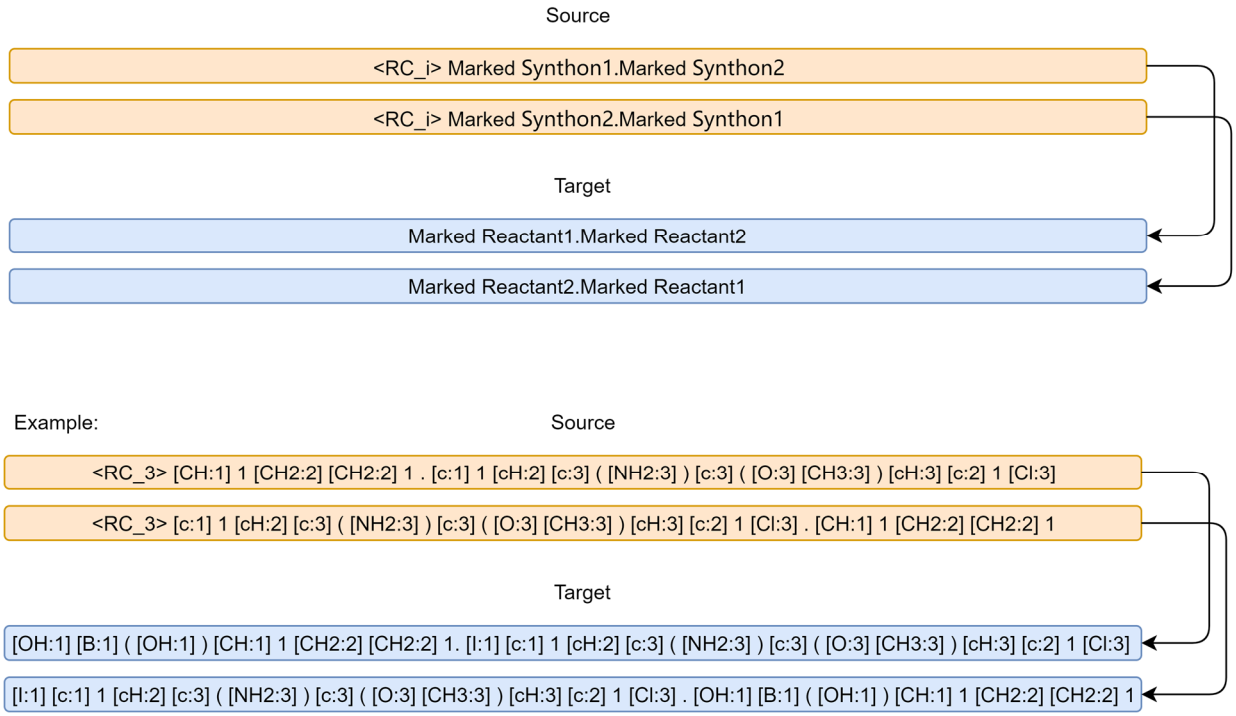

Figure 5. Synthons-to-Reactants datasets augmentation. <RC_i> is the reaction type if applicable.

2.2.3. Synthons-to-Reactants dataset

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 section 2.2.1. 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 Section 1.1) 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[36], and the properly 'labelled' SMILES can be produced with the RDKit[36] *MolToSmiles* function. This is how we prepare the source (input) part of this dataset. As for the corresponding targe (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 4, 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"**.**

2.2.4. Data augmentation for Synthons-to-Reactants dataset

As shown in Figure 4, 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 Figure 5 and Table 3.

2.2.5. 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)[38]. There are 1,808,937 raw records. For reactions involving multiple products, we duplicate the

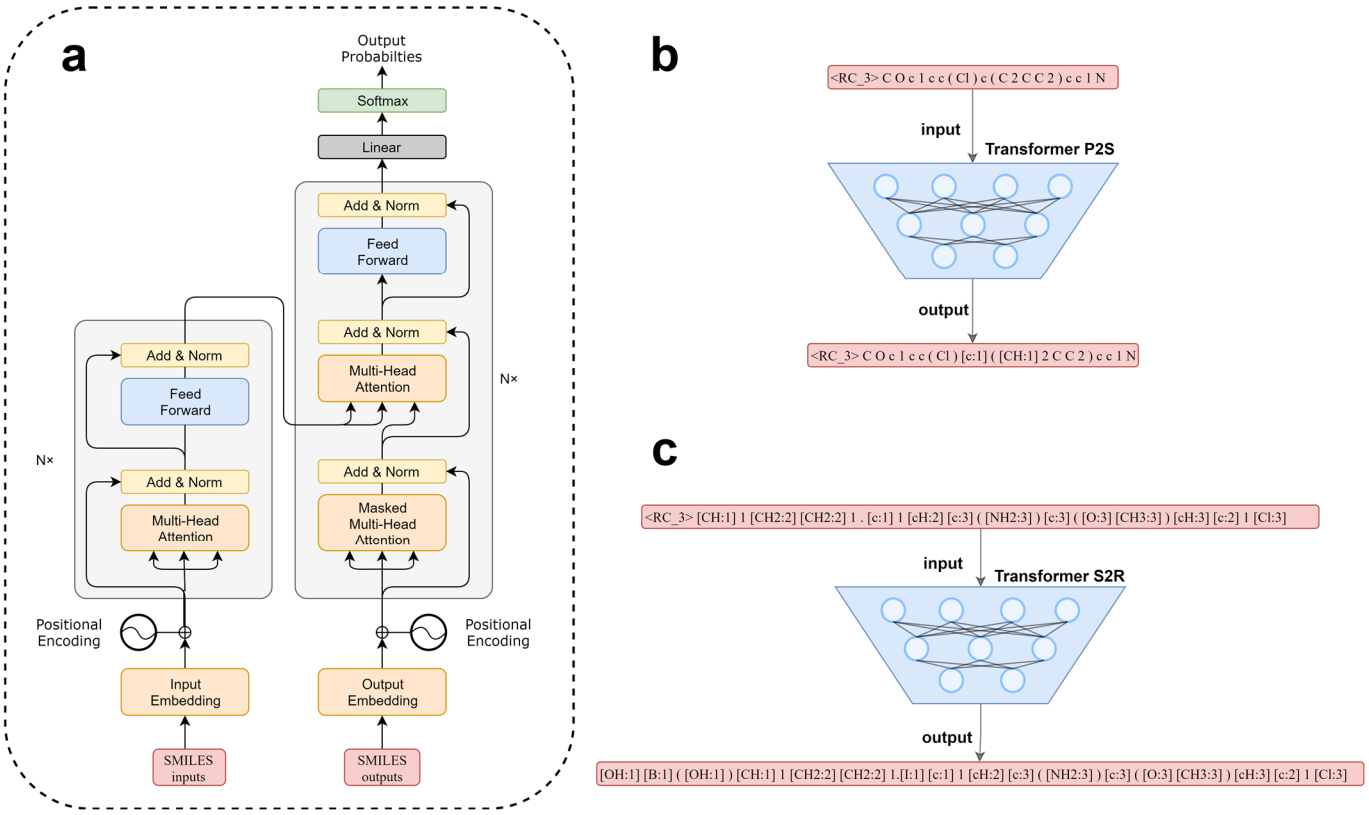

Figure 6. Models and their input and output. The parameters of the two models are the same except for the word embedding layer. (a) Transformer P2S and S2R model architecture. (b) Both the input and output of Transformer P2S use canonical SMILES and have reaction type tokens if applicable. (c) Atom symbol 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.

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 Section 2.2.1 to 2.2.4, 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 Section 1.2 and Section 1.4, respectively. Further details of these large-scale datasets are summarized in Table 1,2,3

2.3. Models training

As mentioned earlier, we utilize two transformer[31] models in the proposed workflow. Both are built with Open NMT[39]. 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 6. See **Table S1** for further details such as the hyperparameters and training details in the Supporting Information.

2.4. 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

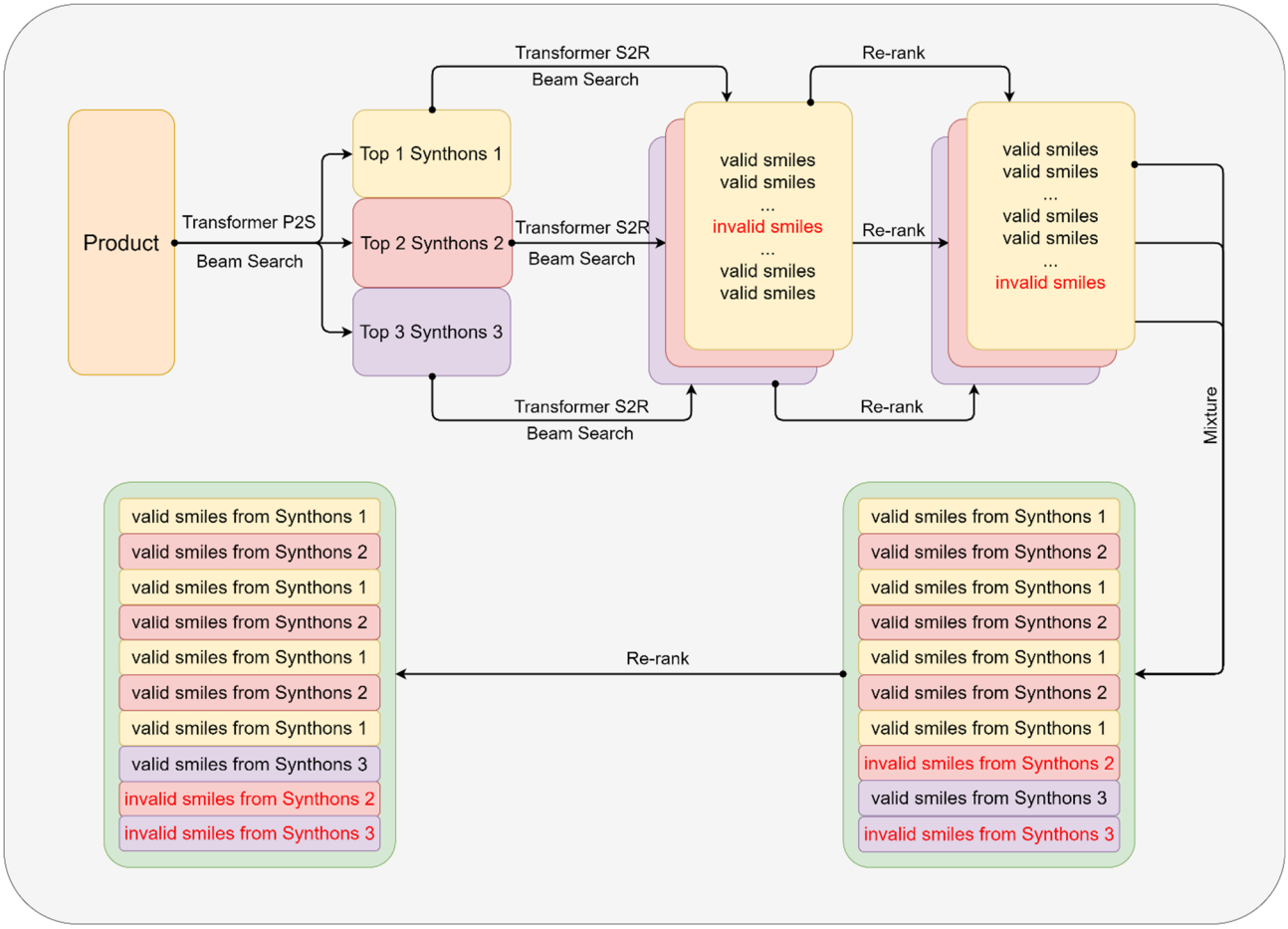

Figure 7. Mix and Match. We select the rank 1-3 synthons predicted by Transformer P2S and send them to Transformer S2R to predict the reactants, and the obtained results are alternately combined. Use the re-rank approach to rank invalid SMILES at the end.

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 dataset.

2.5. Reaction Diversity

For reactions with unknown reaction type, we check whether RetroPrime can offer diverse reaction outcomes. We use a reaction type predictor[33] based on typical message-passing graph convolution network[40] 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.

2.6. 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 7. See Supporting Information Figure S1 for further details on the "mix-and-match".

2.7 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 Section 2.2.3. 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.

3. Results and Discussion

3.1. Baseline

We benchmark our method against six baselines, including four template-free and two template-based methods. Specifically, Seq2Seq[19] is a template-free approach that trains an LSTM model to translate the SMILES of target molecules to SMILES of reactants. Similarity 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. Molecular Transformer[30] is similar to the Seq2Seq translation model but using Transformer instead of LSTM architecture at core. G2Gs[20] and GraphRetro[22] 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.

3.2. Top-N accuracy

For the USPTO-50K dataset, our results are presented in Table 4 and Table 5, 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 6, 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.

Table 4. 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[19]	37.4	52.4	57.0	61.7
Coley et al. RetroSim[16]	52.9	73.8	81.2	88.1
Molecular Transformer[30]	57.3	71.6	75.2	78.0
Shi et al. G2Gs[20]	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[22]	**67.8**	**82.7**	85.3	87.0

Table 5. 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
Molecular 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[4]	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 6. 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 7. Reaction type analysis on USPTO-50K dataset when the reaction type is unknown.

Methods	Reaction type/product
Molecular Transformer	1.74
RetroPrime	**2.40**

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 Molecular Transformer (as both approaches mainly use Transformer to make predictions). This diversity estimation, based on the second metrics introduced in Section 2.5, is shown in Table 7. 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 Molecular Transformer. As show in Figure 8, RetroPrime generates more diverse results, comparing to the baseline models. This diversity comes from the 'Mix and Match' strategy described in Section 2.6. Additional results are provided in Supporting Information.

3.3. 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 8 and 9 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.

Table 8. 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 9. 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

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

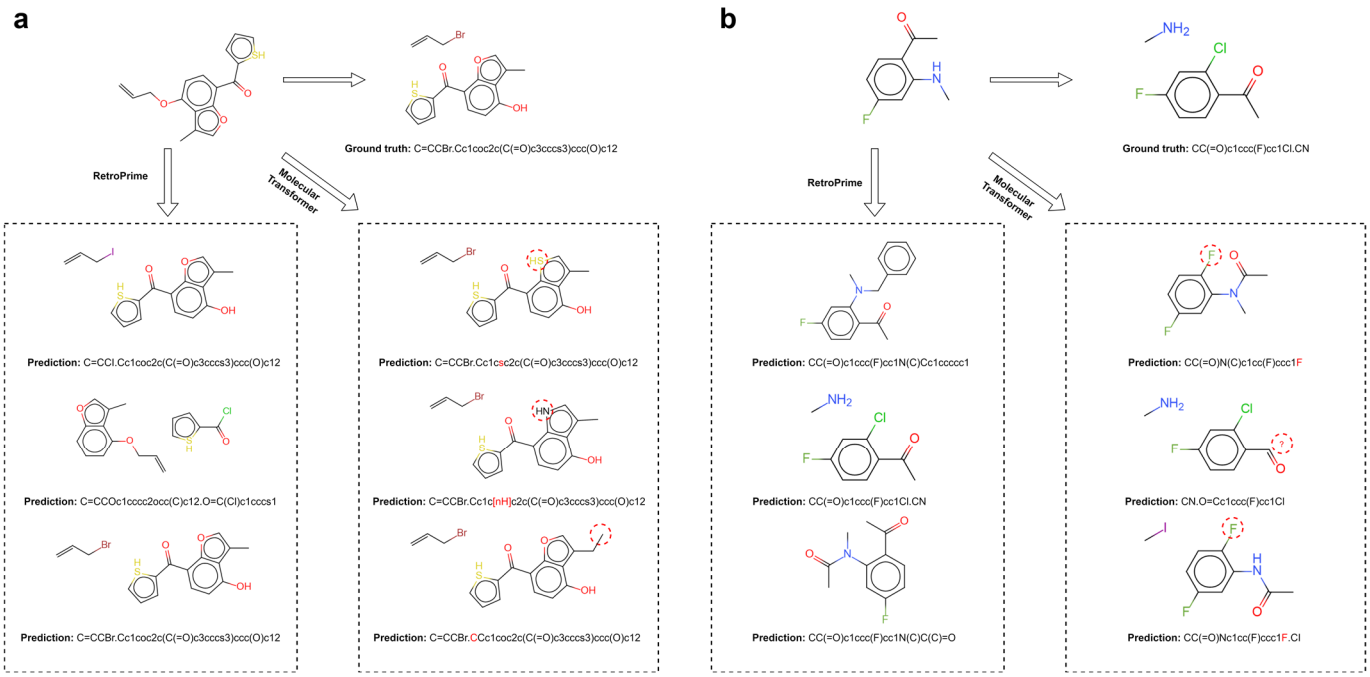

Figure 8. Comparing the predictions of RetroPrime and Molecular Transformer when the reaction type is unknown, we can see that RetroPrime can predict a variety of different retrosynthesis schemes while the results generated by the Molecular Transformer are similar. While many results generated by the Molecular Transformer are very close to ground truth, but the overall reaction is not chemically plausible.

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.[34]. In short, we feed the predictions results (i.e. reactants) of our retrosynthetic method to a forward reaction prediction model, the Molecular Transformer[32]. 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%[32] for the Top-5 accuracy.

Results of this scrutiny on chemical validity of our method are summarized in Table 10 and 11. 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.

Table 10. 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
S2R Marked smiles	50060	48053	**45.2**	**53.5**	**55.6**
S2R Canonical smiles	50060	48637	33.7	40.4	42.3
Molecular Transformer	50060	47121	32.9	39.5	41.4

Table 11. 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
S2R Marked smiles	50060	49786	**46.8**	**55.9**	**58.3**
S2R Canonical smiles	50060	49790	42.0	49.7	51.8
Molecular Transformer	50060	48004	36.4	43.7	45.8

3.4 Generalizability across chemical space

Table 12. 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
Molecular Transformer	21.9	30.5	33.0	34.6
RetroPrime	**24.4**	**34.8**	**37.2**	**40.7**

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[16] and Molecular 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 training set. The results are shown in Table 12. In principle, USPTO-full contains a lot more molecules that are not similar to the ones in USPTO-50K. The results in Table 12 show that RetroPrime exhibits better generalizability. While Molecular 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.

4. 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 3.4 further highlight RetroPrime's generalizability across chemical space. These encouraging results seems to concur with an earlier

observation[41] 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 Retro-Prime 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.

AUTHOR INFORMATION

Corresponding Author

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

***E-mail**: kimhsieh@tencent.com.

***E-mail**: bliao@tencent.com.

Notes

The authors declare no competing financial interest.

Acknowledgements

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

References

(1) Corey, E. J. The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture). *Angew. Chemie Int. Ed. English* **1991**, *30* (5), 455–465. https://doi.org/10.1002/anie.199104553.

(2) Ott, M. A.; Noordik, J. H. Computer Tools for Reaction Retrieval and Synthesis Planning in Organic Chemistry. A Brief Review of Their History, Methods, and Programs. *Recl. des Trav. Chim. des Pays-Bas* **1992**, *111* (6), 239–246. https://doi.org/10.1002/recl.19921110601.

(3) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34* (3), 247–266.

(4) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. **2020**, No. NeurIPS, 1–15.

(5) 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* (1), 79–107.

(6) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inform.* **2014**, *33* (6–7), 469–476. https://doi.org/10.1002/minf.201400052.

(7) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal

Chemistry Synthesis. *J. Med. Chem.* **2020**. https://doi.org/10.1021/acs.jmedchem.9b02120.

(8) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14* (1), 19–38. https://doi.org/10.1351/pac196714010019.

(9) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chemie (International Ed. English)* **1996**, *34* (23–24), 2613–2633. https://doi.org/10.1002/anie.199526131.

(10) Engkvist, O.; Norrby, P. O.; Selmi, N.; Lam, Y. hong; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discov. Today* **2018**, *23* (6), 1203–1218. https://doi.org/10.1016/j.drudis.2018.02.014.

(11) Feng, F.; Lai, L.; Pei, J. Computational Chemical Synthesis Analysis and Pathway Design. *Front. Chem.* **2018**, *6* (JUN), 199. https://doi.org/10.3389/fchem.2018.00199.

(12) Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chemie - Int. Ed.* **2015**, *54* (11), 3449–3464. https://doi.org/10.1002/anie.201410744.

(13) Caramelli, D.; Granda, J. M.; Cambié, D.; Mehr, S. H. M.; Henson, A. An Artificial Intelligence That Discovers Unpredictable Chemical Reactions.

(14) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving Laboratories. *Trends Chem.* **2019**, *1* (3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007.

(15) Nair, V. H.; Schwaller, P.; Laino, T. Data-Driven Chemical Reaction Prediction and Retrosynthesis. *Chimia (Aarau).* **2019**, *73* (12), 997–1000. https://doi.org/10.2533/chimia.2019.997.

(16) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3* (12), 1237–1245. https://doi.org/10.1021/acscentsci.7b00355.

(17) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - A Eur. J.* **2017**, *23* (25), 5966–5971. https://doi.org/10.1002/chem.201605499.

(18) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic Retrosynthetic Route Planning Using Template-Free Models. *Chem. Sci.* **2020**, *11* (12), 3355–3364. https://doi.org/10.1039/c9sc03666k.

(19) 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* (10), 1103–1113. https://doi.org/10.1021/acscentsci.7b00303.

(20) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. *arXiv Prepr. arXiv2003.12725* **2020**.

(21) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv Prepr. arXiv1612.09529* **2016**.

(22) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learning Graph Models for Template-Free Retrosynthesis. *arXiv Prepr. arXiv2006.07038* **2020**.

(23) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59* (6), 2529–2537. https://doi.org/10.1021/acs.jcim.9b00286.

(24) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Knew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602. https://doi.org/10.1021/ci800228y.

(25) Weininger, D. SMILES, a Chemical Language and Information System: 1: Introduction to

Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

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

(27) Chen, B.; Shen, T.; Jaakkola, T. S.; Barzilay, R. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis. *arXiv Prepr. arXiv1910.09688* **2019**.

(28) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (1), 47–55. https://doi.org/10.1021/acs.jcim.9b00949.

(29) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis. 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. https://doi.org/10.1007/978-3-030-30493-5_78.

(30) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-Mcleod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155. https://doi.org/10.1039/c9cc05122h.

(31) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; 2017; Vol. 2017-Decem, pp 5999–6009.

(32) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

(33) Chen, B.; Shen, T.; Jaakkola, T. S.; Barzilay, R. Learning to Make Generalizable and Diverse Predictions for Retrosynthesis. **2019**, 1–11.

(34) Schwaller, P.; Petraglia, R.; Laino, T. Evaluation Metrics for Single-Step Retrosynthetic Models. **2019**, No. NeurIPS.

(35) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56* (12), 2336–2346. https://doi.org/10.1021/acs.jcim.6b00564.

(36) Landrum, G. RDKit: Open-Source Cheminformatics. 2006.

(37) Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv Prepr. arXiv1703.07076* **2017**.

(38) Lowe, D. Chemical Reactions from US Patents (1976-Sep2016). 2018.

(39) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. M. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ACL 2017 - 55th Annu. Meet. Assoc. Comput. Linguist. Proc. Syst. Demonstr.* **2017**, 67–72. https://doi.org/10.18653/v1/P17-4012.

(40) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *Adv. Neural Inf. Process. Syst.* **2017**, *2017-Decem* (Nips), 2608–2617.

(41) Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155.