Interpretable Retrosynthesis Prediction in Two Steps

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, Junzhou Huang

Retrosynthesis is the process of recursively decomposing target molecules into available building blocks. Computer-aided retrosynthesis can potentially assist chemists in designing synthetic routes; however, at present, it is cumbersome and lacks interpretability behind their predictions. In this study, we devise a novel template-free model for retrosynthetic expansion by automating the procedure that chemistsusedtodo. Our method plans synthesis in two steps, by first identifying the potential disconnection bonds of the molecule graph with a graph neural network and thereafter generating synthons according to the identified disconnection bonds of the target molecule graph, and then predicting the associated reactants SMILES based on the obtained synthons with a reactant prediction model. While outperforming previous state-of-the-art baselines by a significant margin on the benchmark datasets, our model also provides predictions with high diversity and chemically reasonable interpretation.

File list (1)

Interpretable Retrosynthesis Prediction in Two Steps.pdf (475.69 KiB)   view on ChemRxiv • download file

INTERPRETABLE RETROSYNTHESIS PREDICTION IN TWO STEPS

Chaochao Yan[*†]
University of Texas at Arlington
chaochao.yan@mavs.uta.edu

Qianggang Ding [*†]
Tsinghua University
dqg18@mails.tsinghua.edu.cn

Peilin Zhao
Tencent AI Lab

Shuangjia Zheng [†]
Sun Yat-sen University

Jinyu Yang [†]
University of Texas at Arlington

Yang Yu
Tencent AI Lab

Junzhou Huang
University of Texas at Arlington
Tencent AI Lab

ABSTRACT

Retrosynthesis is the process of recursively decomposing target molecules into available building blocks. Computer-aided retrosynthesis can potentially assist chemists in designing synthetic routes; however, at present, it is cumbersome and lacks interpretability behind their predictions. In this study, we devise a novel template-free model for retrosynthetic expansion by automating the procedure that chemists used to do. Our method plans synthesis in two steps, by first identifying the potential disconnection bonds of the molecule graph with a graph neural network and thereafter generating synthons according to the identified disconnection bonds of the target molecule graph, and then predicting the associated reactants SMILES based on the obtained synthons with a reactant prediction model. While outperforming previous state-of-the-art baselines by a significant margin on the benchmark datasets, our model also provides predictions with high diversity and chemically reasonable interpretation.

1   Introduction

Retrosynthesis to a desired compound is commonly constructed by recursively decomposing it into a set of available reaction building blocks. This analysis mode was formalized in the pioneering work [1] [2] and now becomes one of the fundamental modules of modern chemical society. This process is challenging, in part due to the size of the search space and incomplete understanding of the reaction mechanism. Typically, retrosynthesis is undertaken by human experts, and therefore it is a subjective process and requires the expertise and experience.

The formulation of retrosynthesis can be decomposed into two sub-tasks [1] [3]: (i) *Breaking down* the given target molecule into a set of corrupted units, which are also known as synthons. (ii) *Calibrating* the predicted synthons into a set of reactants, each of which corresponds to an available molecule. Since a molecule may has lots of possible retrosynthetic routes, it is challenging to select the most appropriate one since the feasibility of a route is often determined by the global chemical structure of the molecule.

To this end, many computational methods [4] [5] [6] [7] [8] [9] have been developed to assist in designing synthetic routes for novel molecules, among which most are dependent on hand-coded reaction templates. However, it is not practical to manually encode all the synthesis routes considering the exponential growth in the number of published reactions. Recently, some template-based methods [6] [8] [9] propose to automatically extract reaction templates from the reaction databases and choose appropriate rules for the selected templates to apply to the target molecules. The key process of these approaches is to select relevant templates for target molecules. However, these methods rely heavily on the predefined atom dictionary to map atoms between reactants and products, which is still a nontrivial problem [10] [11]. Furthermore, commonly used tools to identify the atom-mapping are based on databases of expert rules and templates, which seems to get stuck in an infinite loop [12]. What is worse, template-based models can only infer

---

[*]Both authors contribute equally to the work.

[†]This work is done when Chaochao Yan, Qianggang Ding, Shuangjia Zheng, and Jinyu Yang work as interns at Tencent AI Lab.

reactions within the chemical space covered by the template's database, and they are restrained from discovering novel chemistry [13].

On the other hand, some research [7] [14] also attempts to use a SMILES sequence representation for the reactants and products, and they simply treat the retrosynthesis as a machine translation problem. Although simple and expressive, these approaches ignore rich chemistry knowledge within the chemical reaction and lack of generalization. Besides, such models do not fit into the chemists' analytical process and lack interpretability behind their predictions.

To overcome these challenges, inspired by the expert experience from chemists, we devise a two-step framework to automate the procedure that chemists used to do for retrosynthesis prediction. Note that in this work we focus on the "single step" version (predict possible reactants given the target product) of retrosynthesis problem following previous methods [7] [9] [14]. Our model tackles the "single step" retrosynthesis prediction problem in two steps. **Firstly**, we propose to identify the potential disconnection edges (bonds) of the target molecule with a novel Edge-informed Graph Attention Network (EGAT), which considers the molecule chemical knowledge by taking atom features and chemical bond features as the graph node and edge features, respectively. Note that this step can also be considered as predicting the reaction center, which chemists think is the key for retrosynthesis prediction. **Secondly**, once the synthons are obtained by splitting the target molecule according to the identified disconnection bonds, our Reactant Prediction Network (RPN) will try to predict associated reactants given the target and synthons accordingly. Unlike in previous methods [7] [14], the reactants generation order is undetermined since there is no such supervised correspondence information in the training data, which makes the model training converge much slower and inferior, the reactants generation order is unique in our method. Besides, we find the robustness of the RPN plays an important role in the retrosynthesis. We propose to augment the training data of the RPN by incorporating unsuccessful synthon predictions, which make our RPN able to correct the error and make successful reactant predictions.

Our main contribution can be summarized as follows:

(1) We propose to identify the potential disconnection edges with a novel Edge-informed Graph Attention Network (EGAT) which is largely dependent on the chemical knowledge.

(2) By splitting the target molecule into synthons, our RPN is able to determine the reactants generation order. We further propose to augment the training data by introducing unsuccessfully predicted synthons, and which makes our RPN extreme robust and achieve significant improvements.

(3) On the standard USPTO-50K dataset, our proposed method achieves 70.4% and 65.5% Top-1 retrosynthesis accuracy when w/ and wo/ given reaction types, respectively , which outperforms SOTA accuracy 63.2% (w/) and 52.6% (wo/) reported in [9] by a large margin.

2 Methodology

Notations.   Given a molecule $\mathbf{M}$ with N atoms, let $\mathbf{S}$ be the SMILES [15] representation and $\mathbf{G}$ be the graphical representation for the molecule, respectively. For the molecule graph $\mathbf{G}$, denotes the matrix representation of the molecule atom features as $X \in \mathbb{R}^{N \times F}$, the tensor representation of the molecule bond features as $E \in \mathbb{R}^{N \times N \times D}$, and the adjacency matrix of $\mathbf{G}$ as $A \in \mathbb{R}^{N \times N \times D}$. Specifically, let $x_i \in \mathbb{R}^F, i = 1, 2, 3, ..., N$ be the F dimensional feature vector of the atom (node) i. Similarly, $e_{ij} \in \mathbb{R}^D, i = 1, 2, 3, ..., N; j = 1, 2, 3, ..., N$ be the D dimensional feature vector of the bond connecting the atom i and the atom j if the bond $b_{i,j}$ exists. We use P, S, R to presents the product, synthons, and reactants in the reaction formulation, respectively.

Problem Formulation.   The one-step retrosynthesis prediction problem $(P \rightarrow R)$ can be described as given the desired product P, finding out a set of reactants $R = \{R_1, R_2, ...\}$ that can generate the major product P through a valid chemical reaction. It can be regarded as the reverse process of the forward reaction prediction problem $(R \rightarrow P)$, which predicts the outcome product P given a set of reactants $R = \{R_1, R_2, ...\}$.

Method Overview.   Before diving into the details of our proposed retrosynthesis prediction framework, let us first loot at the blueprint of our proposed model as illustrated in Figure 1. Briefly, we can divide the overall retrosynthesis task $(P \rightarrow R)$ into two closely dependent steps **bond disconnection** $(P \rightarrow S)$ and **reactant prediction** $(S \rightarrow R)$. The first step bond disconnection is to predict the potential reaction bonds which are changed during the reaction, then we can break down the product P into a set of intermediates called synthons $S = \{S_1, S_2, ...\}$. Note that each synthon S_i corresponds to a reactant molecule R_i. The second step is to calibrate and transform the predicted synthons $S = \{S_1, S_2, ...\}$ to the associated reactants $R = \{R_1, R_2, ...\}$. Although the intermediate synthons are not needed, decomposing the original retrosynthesis task $(P \rightarrow R)$ into two steps $P \rightarrow S$ and $S \rightarrow R$ can have multiple benefits which will be elaborated thoroughly in the following paragraphs.
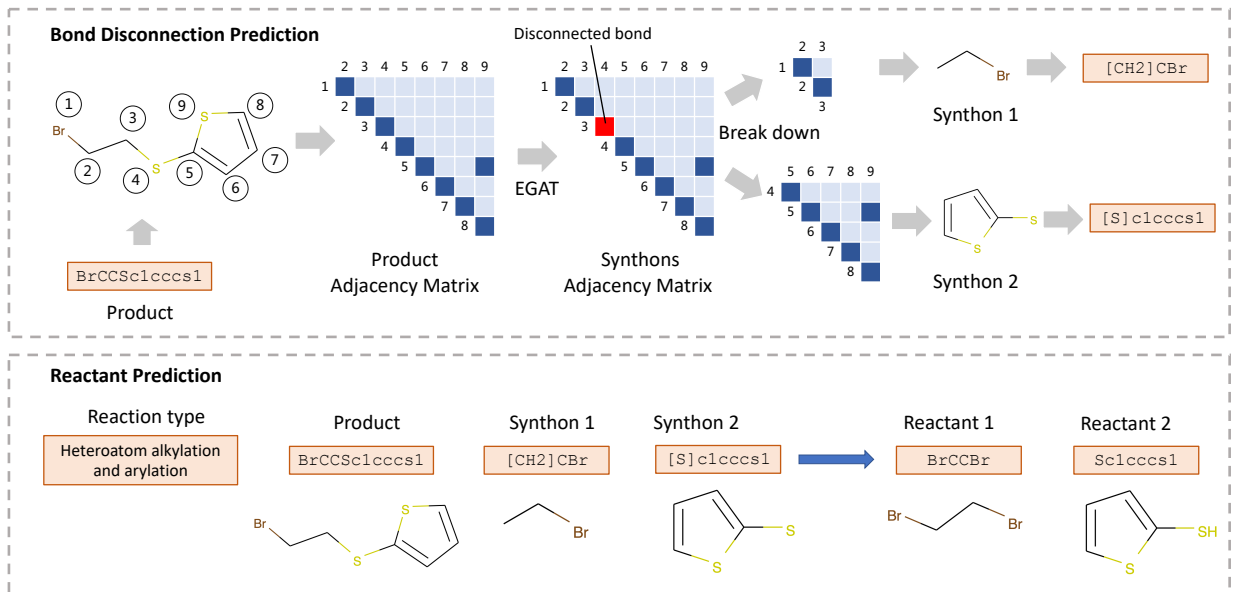
Figure 1: Pipeline overview. We conduct retrosynthesis in two dependent steps **bond disconnection prediction** and **reactant prediction**. The first step takes as input the molecule graph built from the product canonical SMILES, and generates intermediate synthons according to the disconnection prediction results. The second step is to generate reactants corresponding to the synthons based on the given product, synthons, and reaction type if applicable.

2.1 EGAT for Bond Disconnection Prediction

We treat the bond disconnection prediction as a graph-to-graph transformation problem which is similar to the forward reaction outcome prediction [16]. Given the canonical SMILES representation S_P for a product molecule, we firstly convert the SMILES string S_P to the corresponding molecule graph G_P with N_P nodes. We propose a graph neural network named Edge-informed Graph Attention Network (EGAT) which takes the molecule graph G_P as the input and predicts the bond disconnection within the graph, which is the main task. What is more, we add an auxiliary graph-level task to restrain the total number of disconnection bonds since there may be multiple combinations of disconnection bonds for a given product. Note that all used features and labels can be extracted automatically from the retrosynthesis dataset.
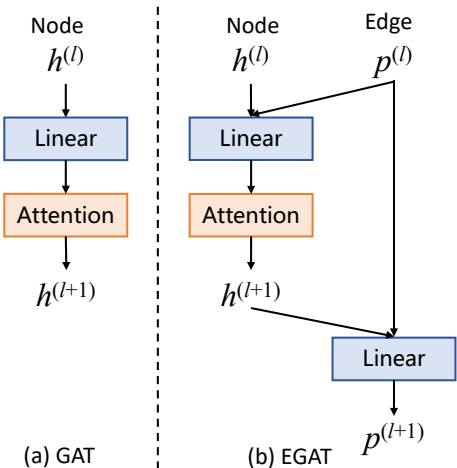

Figure 2: (a) GAT and (b) EGAT.

Distinct from the traditional Graph Attention Network (GAT) [17] which is designed to learn node or graph-level representation, our proposed EGAT is also an edge-based model predicting which bonds should be disconnected in

the retrosynthesis. As illustrated in Figure 2, given the input $\mathbf{G}_p = \{A_p, E_p, X_p\}$, different from the GAT which only considers the node embedding $h_i^{(l+1)}$, our EGAT computes both the node embedding $h_i^{(l+1)}$ and the edge embedding $p_{i,j}^{(l+1)}$ of the layer $l + 1$ from the previous layer's embedding $h_i^{(l)}$ and $p_{i,j}^{(l)}$ by the following equations:

$$z_i^{(l)} = \mathbf{W}^{(l)} h_i^{(l)}, \tag{1}$$

$$c_{i,j}^{(l)} = \text{LeakyReLU}(\mathbf{a}^{(l)^T}[z_i^{(l)}||z_j^{(l)}||p_{i,j}^{(l)}]), \tag{2}$$

$$\alpha_{i,j}^{(l)} = \frac{\exp(c_{i,j}^{(l)})}{\sum_{k \in \mathcal{N}_i} \exp(c_{i,k}^{(l)})}, \tag{3}$$

$$h_i^{(l+1)} = \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{i,j}^{(l)} \mathbf{U}^{(l)}[z_j^{(l)}||p_{i,j}^{(l)}]\Big), \tag{4}$$

$$p_{i,j}^{(l+1)} = \mathbf{V}^{(l)}[h_i^{(l+1)}||h_j^{(l+1)}||p_{i,j}^{(l)}], \tag{5}$$

where the initial embeddings $h_i^{(0)}, p_{i,j}^{(0)}$ are the input node and edge features $x_i, e_{i,j}$, $\mathbf{W}^{(l)} \in \mathbb{R}^{F'^{(l)} \times F^{(l)}}$, $\mathbf{a}^{(l)} \in \mathbb{R}^{2F'^{(l)}+D^{(l)}}$, $\mathbf{U}^{(l)} \in \mathbb{R}^{F^{(l+1)} \times (F'^{(l)}+D^{(l)})}$, $\mathbf{V}^{(l)} \in \mathbb{R}^{D^{(l+1)} \times (2F^{(l+1)}+D^{(l)})}$ are trainable parameters in the EGAT, $\mathcal{N}_i$ is the neighbor nodes of the node i, $\alpha_{i,j}$ is the attention weight between the node i and its neighbor node j, and $h_i^{(l+1)} \in \mathbb{R}^{F^{(l+1)}}$ as well as $p_{i,j}^{(l+1)} \in \mathbb{R}^{D^{(l+1)}}$ are the output node and and edge representations, respectively.

After stacking L EGAT layers, we can obtain the final edge representation $p_{i,j}^{(L)}$ for the chemical bond between node i and j, as well as node representation $h_i^{(L)}$ for each node i. Note that we can also apply multi-head attention mechanism like the original GAT. To predict the disconnection probability for the bond $p_{i,j}^{(L)}$, we perform a fully-connected layer parameterized with $\mathbf{w}_{fc} \in \mathbb{R}^{D^{(L)}}$ and a $Sigmoid$ activation layer to the $p_{i,j}^{(L)}$:

$$d_{i,j} = \text{Sigmoid}(\mathbf{w}_{fc}^T \cdot p_{i,j}^{(L)}). \tag{6}$$

Our optimization goal for bond disconnection prediction is to minimize the negative log-likelihood between the disconnection prediction $d_{i,j}$ and the ground-truth $y_{i,j} \in \{0, 1\}$ through the following binary cross entropy loss function:

$$\mathcal{L}_{\text{EGAT}} = -\frac{1}{K} \sum_{k=1}^{K} \sum_{b_{i,j} \in \mathbf{G}_k} [(1 - y_{i,j}) \log(1 - d_{i,j}) + y_{i,j} \log(d_{i,j})], \tag{7}$$

where K is the total data size and bond $b_{i,j}$ exists if the corresponding adjacency element $a_{i,j}$ is non-zero.

Molecule Atom and Bond Features. The atom and bond features are listed in Tables 1 and 2, respectively. The most of them are the same to that used in [18] which is for chemical property prediction. We compute these features using the open-source package RDKit [19]. To utilize the rich chemical information, we add a functional groups indicator to the atom feature. Functional groups are automatically extracted from the product SMILES. Besides, for the retrosynthesis dataset with known reaction types, a reaction type indicator is also added to the atom features.

2.2 Reactant Prediction Network

Once we have the bond disconnection prediction results then the corresponding synthons can be obtained immediately by applying the bond disconnection to split the product molecule. We convert the split synthon graphs to SMILES representations using the RDKit tool, though these synthons may be invalid molecule structures. Since each synthon is basically a sub-structure of a reactant, now the remained task $S \rightarrow R$ is much simpler than the original problem $P \rightarrow R$ in which even the number of reactants is unknown.

Motivated by the success of the Transformer [20] in natural machine translation, the Transformer is also adapted for reaction outcome prediction [21] and retrosynthesis [14], in which both the product and reactant are represented in SMILES. Transformer is a sequence-to-sequence model equipped with attention mechanism, both the self-attention and the "encoder-decoder" attention [20]. For more details about the Transformer, we refer readers to [20] and The Illustrated Transformer.

We represent molecules in SMILES, and build a reactant prediction network based on the Transformer module. Different from previous methods treating the Transformer as a magic black box, we inject more chemical knowledge to make the

Table 1: Atom Features used in EGAT. All features are one-hot encoding, except the atomic mass is a real number scaled to be on the same order of magnitude.

Feature	Description	Size
Functional groups	All functional groups the atom may belong to.	791
Reaction type	The specified reaction type if it exists.	10
Atom type	Type of atom (ex. C, N, O), by atomic number.	100
# Bonds	Number of bonds the atom is involved in.	6
Formal charge	Integer electronic charge assigned to atom.	5
Chirality	Unspecified, tetrahedral CW/CCW, or other.	4
# Hs	Number of bonded Hydrogen atom.	5
Hybridization	sp, sp2, sp3, sp3d, or sp3d2.	5
Aromaticity	Whether this atom is part of an aromatic system.	1
Atomic mass	Mass of the atom, divided by 100.	1

Table 2: Bond Features used in EGAT. All features are one-hot encoding.

Feature	Description	Size
Bond type	Single, double, triple, or aromatic.	4
Conjugated	Whether the bond is conjugated.	1
In ring	Whether the bond is part of a ring.	1
Stereo	None, any, E/Z or cis/trans.	6

Source	Target
\<RXN_k\> Product \<LINK\> Synthon1.Synthon2	⟹ Reactent1.Reactent2
\<RXN_k\> Product \<LINK\> Synthon2.Synthon1	⟹ Reactent2.Reactent1

Example:

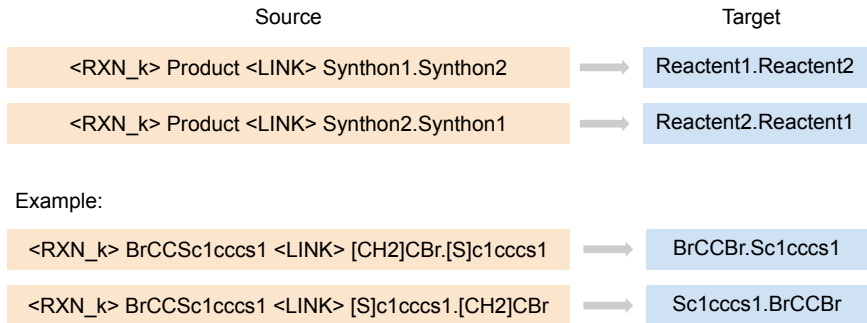| Source	Target
\<RXN_k\> BrCCSc1cccs1 \<LINK\> [CH2]CBr.[S]c1cccs1	⟹ BrCCBr.Sc1cccs1
\<RXN_k\> BrCCSc1cccs1 \<LINK\> [S]c1cccs1.[CH2]CBr	⟹ Sc1cccs1.BrCCBr

Figure 3: Source and target sequences arrangement. \<RXN_k\> is the reaction type if applicable, and product SMILES and synthons are concatenated with a special \<LINK\> token. Multiple synthons are concatenated with a dot following the RDKit.

prediction more reasonable, interpretable, and the most importantly more accurate. Specifically, as shown in Figure 3, the source sequence data is the concatenation of reaction type (if applicable), canonical SMILES of the product, and the associated synthons. The target sequence is the desired reactants arranged according to the synthons. Since the retrosynthesis is closely related to the given product, so the product SMILES is also input into the RPN encoder. What is more, all synthons should also be observable when predicting the corresponding reactant for each synthon, since the reaction occurs with all reactants. This is achieved through the attention mechanism. With the self-attention mechanism, the product and all synthons are observed when embedding the source sequence, they are also considered when generating the reactant sequence through the "encoder-decoder" attention.

Determine Reactants Order. For the first time, the reactants generation order (e.g. which reactant to be generated first if there are two) can be determined by aligning reactants in the target with the synthons in the source, thanks to the intermediate synthons which are associated with reactants one to one. The generation order of reactants is undetermined in previous methods [7] [14], which naively treats the sequence-to-sequence model as a black box. This uncertainty makes the model training more difficult since there may be multiple orders for the similar training samples. Our method can eliminate the uncertainty easily. Besides, since the synthons order may also be different between the train and

5

validation/test sets, we augment the training samples with reversely arranged synthons and reactants as shown in Figure 3. We find it works pretty well in our experiments and the comprehensive permutations of synthons is unnecessary.

Make the RPN Robust. To make our RPN robust enough and able to predict reactants even if the disconnection bond prediction is different from the given ground truth, we further augment the RPN training set by including those unsuccessfully predicted synthons on training data. We do not reverse the order of synthons for these augmentation samples like in Figure 3. We find this measure very effective and will show the improvement in experimental sections.

3 Experiments

3.1 Datasets

We experiment our method on the standard retrosynthesis dataset USPTO-50K [22] which is derived from the USPTO granted patents [23]. USPTO-50K is a standard benchmark which contains 50,000 reactions with 10 reaction types overall. It is widely used in previous retrosynthesis work [7] [6] [14] [9]. We adopt the randomly split training, testing, and validation datasets with 40K, 5K, and 5K, respectively, kindly provided by [6]. For the second step RPN training, we add an extra 28K samples of which the synthons are reversed if there are at least two synthons, and there are 68K training samples in total. We still denote it as USPTO-50K in the following text.

3.2 Settings

Metric. For the retrosynthesis, we use the Top-N accuracy as the evaluation metric which is widely used in the previous work. Beam search [24] strategy is adapted to keep the Top-N predictions throughout the generation process. An correct predicted set of reactants must be **exactly** the same as the ground-truth reactants. Since a molecule can be represented by multiple variants of SMILES string, we evaluate the canonical version of the generated SMILES representations with the ground-truth ones. We also evaluate the EGAT prediction accuracy at the molecule level, both the disconnection bonds and the number of disconnections, to evaluate the effectiveness of our proposed EGAT. The bond disconnection accuracy is evaluate at the molecule level, in which it is counted as correct when all disconnection bonds are successfully predicted.

Implementation. We use canonical SMILES representation for both the product and reactants. The SMILES string is tokenized with the regular expression from [25]. Our implementation is based on the PyTorch [26]. Particularly, we use the DGL [27] and OpenNMT [28] to implement our EGAT and RPN models, respectively. As for the EGAT, we stack 3 identical 4-head attentive layers of which the hidden dimension is 128. All embeddings size in EGAT is set to 128, such as $F^{(l)}$, $F'^{(l)}$, and $D^{(l)}$, and the N_{max} is set as 4 for the USPTO-50K dataset. We use a threshold of 0.5 to binarize the bond disconnection predictions.

Models Training. We train the EGAT with the loss function $\mathcal{L}_{\text{EGAT}}$ (7) for 80 epochs. The EGAT parameters are optimized with the Adam [29] with default settings, and the initial learning rate is 0.0005 and it is scheduled to multiply 0.2 every 20 epochs. For the RPN parameters, we train independent models on all dataset variants for $300,000$ iterations until the models are well converged. To make fair comparisons, we keep all experimental settings the same when conduct ablation studies, unless we state it explicitly.

3.3 Bond Disconnection Prediction Results

We first report the EGAT results of the bond disconnection prediction. The EGAT prediction accuracy is 86.0% w/ reaction type and 64.0% wo/ reaction type. The reaction type information plays an important role in the retrosynthesis. The reactions of the same type usually share similar reaction patterns (involved atoms, bonds, and functional groups), it is much easier to predict the disconnection bonds if the reaction type is given.

Table 3: Bond disconnection prediction results on test set.

Method	Accuracy (%)	
	W/ reaction type	Wo/ reaction type
EGAT	86.0	64.0

3.4 Reactant Prediction Results

Fot the evaluation of our RPN, we have two strategies on the formation of the evaluation pipelines. The first one is straight forward, we gather the final bond disconnection results from the first step $P \rightarrow S$, and only send the successful disconnection predictions to the second step $S \rightarrow R$. The other strategy is to directly send the generated synthons to the second step without considering if the disconnection predictions successful or not. This means that our RPN need to be robust enough to be able to predict from the "wrong" given synthons, and to correct the error to some extent. We call the second evaluation strategy as the $direct$.

To strengthen the correction ability of our RPN, we also conduct the first step $P \rightarrow S$ prediction on the 40K training data, and the final accuracy are 89.0% with given reaction types. We can obtain about 4K such correction training samples. So the augmented training data size for the RPN is 72K. Without given reaction types, the final accuracy is 70.0% and there are 12K correction training samples, so the training size for the RPN is 80K. We also train RPN models on the augmented datasets, which are denoted as USPTO-50K-aug. The results are reported in Table 4. Note that we use the ground truth synthons from testing dataset as the input for the RPN evaluation. These results indicates the upper bound of the overall retrosynthesis prediction accuracy of our method.

Table 4: $S \rightarrow R$ prediction results. The augmented datasets (72K when w/ reaction types and 80K when wo/ reaction types) are denoted as USPTO-50K-aug, and the USPTO-50K has 68K samples. Experimental results are based on ground truth synthons as input.

Datasets	Reaction Types	Top-n Accuracy (%)					
		1	3	5	10	20	50
USPTO-50K	W/	72.9	86.5	88.3	89.5	90.4	91.6
USPTO-50K-aug		73.4	86.7	88.5	89.7	90.9	92.1
USPTO-50K	Wo/	71.9	85.7	87.5	88.9	90.0	91.0
USPTO-50K-aug		70.9	84.6	86.4	88.2	89.4	90.6

We evaluate the overall retrosynthesis prediction accuracy by the two aforementioned strategies, and report results in Tables 5 and 6. Our method achieves impressive performance on the unseen testing data. For retrosynthesis results w/ given reaction types in Table 5, all four variants of our proposed method achieve better Top-1 accuracy than the best previously reported accuracy 63.2% [9]. Particularly, the model **Ours-aug-direct** trained on augmented dataset achieves 70.4% Top-1 accuracy when evaluated with the $direct$ strategy, which improves the previous SOTA 63.2% by 7.2%. As for the results wo/ given reaction types in Table 6, the model **Ours-aug-direct** also can improve the best Top-1 accuracy from 52.6% [9] to 65.6%.

Table 5: Results when reaction types are given. aug denotes the model is trained on USPTO-50K-aug dataset, and $direct$ indicates it is evaluated by the $direct$ strategy. The best score in each column is displayed in the bold.

Methods	Top-n Accuracy (%)					
	1	3	5	10	20	50
Expert System [7]	35.4	52.3	59.1	65.1	68.6	69.5
Seq2Seq [7]	37.4	52.4	57.0	61.7	65.9	70.7
RetroSim [6]	52.9	73.8	81.2	88.1	91.8	92.9
NeuralSym [30]	55.3	76.0	81.4	85.1	86.5	86.9
SCROP [14]	59.0	74.8	78.1	81.1	-	-
GLN [9]	63.2	77.5	83.4	**89.1**	**92.1**	**93.2**
Ours	63.5	75.2	76.7	77.6	78.3	79.3
Ours-direct	64.0	75.7	77.5	78.3	79.2	80.2
Ours-aug	63.8	75.1	76.5	77.4	78.4	79.3
Ours-aug-direct	**70.4**	**83.4**	**85.3**	86.8	88.1	89.3

Table 6: Results without reaction types. *aug* denotes the model is trained on USPTO-50K-aug dataset, and *direct* indicates it is evaluated by the *direct* strategy. The best score in each column is displayed in the bold.

Methods	Top-n Accuracy (%)					
	1	3	5	10	20	50
RetroSim [6]	37.3	54.7	63.3	74.1	82.0	85.3
NeuralSym [30]	44.4	65.3	72.4	78.9	82.2	83.1
SCROP [14]	43.7	60.0	65.2	68.7	-	-
GLN [9]	52.6	68.0	75.1	83.1	**88.5**	**92.1**
Ours	48.1	56.3	57.3	58.0	58.7	59.1
Ours-direct	48.4	56.9	57.9	58.8	59.6	60.2
Ours-aug	47.6	56.1	57.0	57.9	58.5	59.2
Ours-aug-direct	**65.6**	**78.7**	**80.8**	**83.3**	84.6	86.0

4 Discussion

From the Table 5 and 6, we can find that our proposed method is better at predicting the Top-5 reactant sets, which indicates the predictions of our method lacks of diversity. This is determined by the nature of our RPN model. While template-based methods [6] [9] can achieve high Top-20 and Top-50 accuracy, since these template-based methods can enumerate different reaction templates to make the prediction. To increase the diversity of our predictions, we may design new strategies to select multiple sets of disconnection bonds for each product, which is hopefully able to improve the Top-50 accuracy. This is left as the feature work.

5 Conclusion

We present a two-step retrosynthesis model, and inject more chemical knowledge into the retrosynthesis prediction, which makes the prediction more interpretable and explainable. Our method achieves 70.4% Top-1 accuracy on the standard USPTO-50K dataset with reaction types given, improving the previously reported best performance 63.2% by a large margin.

References

[1] EJ Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses. *Science*, 166(3902):178–192, 1969.

[2] Elias James Corey. The logic of chemical synthesis: multistep synthesis of complex carbogenic molecules (nobel lecture). *Angewandte Chemie International Edition in English*, 30(5):455–465, 1991.

[3] David A Pensak and EJ Corey. Lhasa—logic and heuristics applied to synthetic analysis. 1977.

[4] Clara D Christ, Matthias Zentgraf, and Jan M Kriegl. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *Journal of chemical information and modeling*, 52(7):1745–1756, 2012.

[5] Anders Bogevig, Hans-Jurgen Federsel, Fernando Huerta, Michael G Hutchings, Hans Kraut, Thomas Langer, Peter Low, Christoph Oppawsky, Tobias Rein, and Heinz Saller. Route design in the 21st century: The ic synth software tool as an idea generator for synthesis prediction. *Organic Process Research & Development*, 19(2):357–368, 2015.

[6] Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.

[7] Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.

[8] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.

[9] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, pages 8870–8880, 2019.

[10] William Lingran Chen, David Z Chen, and Keith T Taylor. Automatic reaction mapping and reaction center detection. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(6):560–593, 2013.

[11] Wojciech Jaworski, Sara Szymkuć, Barbara Mikulak-Klucznik, Krzysztof Piecuch, Tomasz Klucznik, Michał Kaźmierowski, Jan Rydzewski, Anna Gambin, and Bartosz A Grzybowski. Automatic mapping of atoms across both simple and complex chemical reactions. *Nature communications*, 10(1):1–11, 2019.

[12] Nadine Schneider, Daniel M Lowe, Roger A Sayle, and Gregory A Landrum. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of chemical information and modeling*, 55(1):39–53, 2015.

[13] Marwin HS Segler and Mark P Waller. Modelling chemical reasoning to predict and invent reactions. *Chemistry–A European Journal*, 23(25):6118–6128, 2017.

[14] Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 2020.

[15] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[16] Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. In *Advances in Neural Information Processing Systems*, pages 2607–2616, 2017.

[17] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[18] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[19] Greg Landrum et al. Rdkit: Open-source cheminformatics. 2006.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[21] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[22] Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.

[23] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

[24] Christoph Tillmann and Hermann Ney. Word reordering and a dynamic programming beam search algorithm for statistical machine translation. *Computational linguistics*, 29(1):97–133, 2003.

[25] Philippe Schwaller, Theophile Gaudin, David Lanyi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science*, 9(28):6091–6098, 2018.

[26] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017.

[27] Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, et al. Deep graph library: Towards efficient and scalable deep learning on graphs. *arXiv preprint arXiv:1909.01315*, 2019.

[28] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*, 2017.

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[30] Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

Interpretable Retrosynthesis Prediction in Two Steps.pdf (475.69 KiB)   view on ChemRxiv   •   download file