Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

UAlign: Pushing the Limit of Template-free Retrosynthesis Prediction with Unsupervised SMILES Alignment

Kaipeng Zeng
Shanghai Jiao Tong University
Bo Yang
Shanghai Jiao Tong University
Xin Zhao
Shanghai Jiao Tong University
Yu Zhang
Shanghai Jiao Tong University
Fan Nie
Shanghai Jiao Tong University
Xiaokang Yang
Shanghai Jiao Tong University
Yaohui Jin
Shanghai Jiao Tong University
Yanyan Xu
yanyanxu@sjtu.edu.cn

Shanghai Jiao Tong University

Research Article

Keywords: Template-Free Retrosynthesis, Deep Learning, Chemical Reactions, Single-step Retrosynthesis Prediction

Posted Date: April 24th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4292195/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Cheminformatics on July 15th, 2024. See the published version at https://doi.org/10.1186/s13321-024-00877-2.

UAlign: Pushing the Limit of Template-free Retrosynthesis Prediction with Unsupervised SMILES Alignment

Kaipeng Zeng[1], Bo Yang[2], Xin Zhao[1], Yu Zhang[1], Fan Nie[3], Xiaokang Yang[1], Yaohui Jin[1*], Yanyan Xu[1*]

[1*]MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, 200240, Shanghai, China.
[2]Frontiers Science Center for Transformative Molecules (FSCTM), Zhangjiang Institute for Advanced Study, Shanghai Jiao Tong University, Shanghai, 200240, Shanghai, China.
[3]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, 200240, Shanghai, China.

*Corresponding author(s). E-mail(s): jinyh@sjtu.edu.cn; yanyanxu@sjtu.edu.cn;

Abstract

Motivation Retrosynthesis planning poses a formidable challenge in the organic chemical industry, particularly in pharmaceuticals. Single-step retrosynthesis prediction, a crucial step in the planning process, has witnessed a surge in interest in recent years due to advancements in AI for science. Various deep learning-based methods have been proposed for this task in recent years, incorporating diverse levels of additional chemical knowledge dependency.

Results This paper introduces UAlign, a template-free graph-to-sequence pipeline for retrosynthesis prediction. By combining graph neural networks and Transformers, our method can more effectively leverage the inherent graph structure of molecules. Based on the fact that the majority of molecule structures remain unchanged during a chemical reaction, we propose a simple yet effective SMILES alignment technique to facilitate the reuse of unchanged structures for reactant generation. Extensive experiments show that our method substantially outperforms state-of-the-art template-free and semi-template-based approaches. Importantly, our template-free method achieves effectiveness comparable to, or even surpasses, established powerful template-based methods.

Scientific contribution We present a novel graph-to-sequence template-free retrosynthesis prediction pipeline that overcomes the limitations of Transformer-based methods in molecular representation learning and insufficient utilization of chemical information. We propose an unsupervised learning mechanism for establishing product-atom correspondence with reactant SMILES tokens, achieving even better results than supervised SMILES alignment methods. Extensive experiments demonstrate that UAlign significantly outperforms state-of-the-art template-free methods and rivals or surpasses template-based approaches, with up to 5% (top-5) and 5.4% (top-10) increased accuracy over the strongest baseline.

Keywords: Template-Free Retrosynthesis, Deep Learning, Chemical Reactions, Single-step Retrosynthesis Prediction

1 Introduction

Retrosynthesis prediction is a crucial task in organic chemistry, aiding in finding efficient synthetic pathways from target molecules to accessible starting materials. Despite significant advancements in chemical synthesis technology, it still remains a challenge in industries like pharmaceuticals. The extensive search space and the incomplete understanding of chemical reaction mechanisms make retrosynthesis prediction difficult, even for experienced chemists. To address this issue, computer-assisted synthetic planning (CASP) has gained increasing attention in recent years, starting from the seminal work by Corey. This paper focuses on single-step retrosynthesis prediction, which is the fundamental step in CASP. It aims to predict the reactants that can lead to a given product molecule through a single reaction step.

Various deep-learning-based single-step retrosynthesis prediction methods have been proposed in recent years. These methods can be broadly classified into three groups based on their dependency on additional chemical knowledge: template-based, semi-template-based and template-free methods. **Template-based** methods [5, 7, 9, 37] require an extra database of reaction templates. They frame the retrosynthesis prediction as a classification or retrieval problem for reaction templates suitable for the given product molecule to be synthesized. Among these solutions, Retrosim [7] utilizes molecular similarity to rank reaction templates; LocalRetro [5] and GLN [9] use graph neural networks to model the relationship between reaction templates and molecules to predict the most suitable reaction template; RetroKNN [37] further improves upon LocalRetro by addressing the issue of data imbalance using K-nearest neighbors (KNN). Template-based methods have strong interpretability and can accurately predict reactants. However, these methods are often unable to cover all cases and suffer from poor scalability due to limitations imposed by the template database.

To overcome the limitations faced by template-based methods, researchers have turned to generative models. **Semi-template-based** methods incorporate chemical knowledge into generative models with the help of chemical toolkits like RDKit [16], breaking free from the limitations imposed by reaction templates. The key idea of most semi-template-based methods [6, 26, 27, 35, 38] is to first convert the product into synthons based on reaction center identification and then complete the synthons

into reactants. Graph neural networks are commonly used for synthon prediction, followed by leaving group attachment [6, 27], conditional graph generation [26], or SMILES generation [38] for reactant completion. Apart from all above, RetroPrime [35] utilizes two independent Transformers to accomplish synthon prediction and reactant generation as separate tasks.

Semi-template-based methods to a certain extent are more in line with chemical intuition. However, these methods increase the complexity of inference and training as they break down retrosynthesis into two subtasks. Failures in synthon prediction directly affect subsequent reactant completion and overall performance. Besides, methods based on leaving group necessitates an extra leaving group database. This requirement, akin to template-based approaches, imposes limitations on the model's scalability.

As generative models, **Template-free** methods opt to generate reactants directly from the given products. In comparison to generating graph structures, SMILES provides a way to represent molecules as strings. Taking advantage of this, most template-free methods [15, 25, 28, 34, 42] use Transformer models to translate between product SMILES and reactants SMILES. In particular, Graph2SMILES [29] replaces the Transformer encoder with a graph neural network, resulting in a permutation-invariant pipeline. There are also methods [21, 40] formulates the generation of reactants as a series of graph generation or editing operation and solve it auto-regressively. Existing template-free methods generally follows an auto-regressive generation strategy and use beam search for the generation process. Consequently, preserving a level of diversity in the resultant outputs has emerged as a critical consideration for template-free methods [33]. Due to the use of SMILES as input and output, most of template-free methods often overlook the rich topological and chemical bond information present in molecular graphs. Moreover, as reactants molecules need to be generated from scratch, template-free methods frequently suffer from validity issues and fail to leverage an important property of retrosynthesis prediction, i.e., the presence of many common substructures between products and reactants.

In this paper, we focus on the template-free generative approach for retrosynthesis prediction. Existing sequence-to-sequence methods have limitations in extracting robust molecular representations. They overlook the abundance of topological information and chemical bonds, and lack the ability to utilize atom descriptors as rich as those in graph-based methods. Furthermore, template-free methods overlook the fact that the molecular graph topology remains largely unaltered from reactants to products during chemical reactions, as they generate reactants from scratch. While there are methods that attempt to solve this problem using supervised SMILES alignment, they require complex data annotation and impact model training. Given these limitations, the following question naturally arises:

Can we effectively leverage the structural information of product molecules using a much simpler approach?

To address these issues and further enhance template-free methods, we propose a novel graph-to-sequence pipeline called UAlign. Our approach employs a specifically designed graph neural network as an encoder, incorporating information from chemical bonds during message passing to create more powerful embeddings for the decoder. We introduce an unsupervised SMILES alignment mechanism that establishes associations

between product atoms and reactant SMILES tokens which reduces the complexity of SMILES generation and enables the model to focus on learning chemical knowledge. Our model outperforms existing template-free methods by a large margin and demonstrates comparable performance against template-based methods.

2 Methods

We introduce UAlign, a novel single-step retrosynthesis prediction model based on an encoder-decoder architecture, as demonstrated in Fig. 1. It's a fully template-free method without any molecule editing operation using RDKit [16]. We propose a specially designed variant of Graph Attention Network, which incorporates the information of chemical bonds to enhance the capability of capturing the structural characteristics of molecules.

2.1 Preliminary

A molecule can be represented as a graph, denoted by $G = (V, E)$, where V represents the atoms and E represents the chemical bonds. The SMILES representation of a molecule can be obtained by performing a depth-first search (DFS) starting from any arbitrary atom in the molecule graph. Given a molecule graph $G = (V, E)$, we can generate multiple DFS orders and each DFS order corresponds to a SMILES representation of the graph. Denoted the set of all possible DFS orders as $\mathcal{D}(V) \subseteq \mathcal{P}(V)$, $\mathcal{P}(V)$ represents all permutations of the set of atoms V. For each DFS order $O \in \mathcal{D}(V)$, we denoted its corresponding SMILES as $Smiles(G, O)$, which lists all atoms in the order dictated by O. To facilitate our subsequent elaboration, we refer to the position of an atom a in the order O as its rank, denoted as $rank(a, O)$. The atom with the minimal rank given order O is then defined as the root atom, denoted as $root(G, O)$.

2.2 EGAT$^+$

Chemical bonds play a significant role in determining the properties of molecules and contain valuable information. Previous studies [12, 20, 39] have demonstrated that incorporating edge information into graph neural networks can greatly enhance their ability to represent molecular structures. To fully leverage the information brought by chemical bonds, we propose a modified version of the Graph Attention Network (GAT) [32] called EGAT$^+$.

Our proposed model explicitly incorporates edge features, which represent the information derived from chemical bonds, into the message passing process. During each iteration of message passing, the EGAT$^+$ applies self-attention to each node and its one-hop neighbors to calculate attention coefficient according to both node features and edge features. It then aggregates the both node and edge features of these neighbors, considering the attention coefficients, to update the node features. Denote the node feature of atom u as $h_u^{(k)}$ and the edge feature between atom u and v as $e_{u,v}^{(k)}$ at k-th iteration of message passing. In math, the message passing mechanism can be

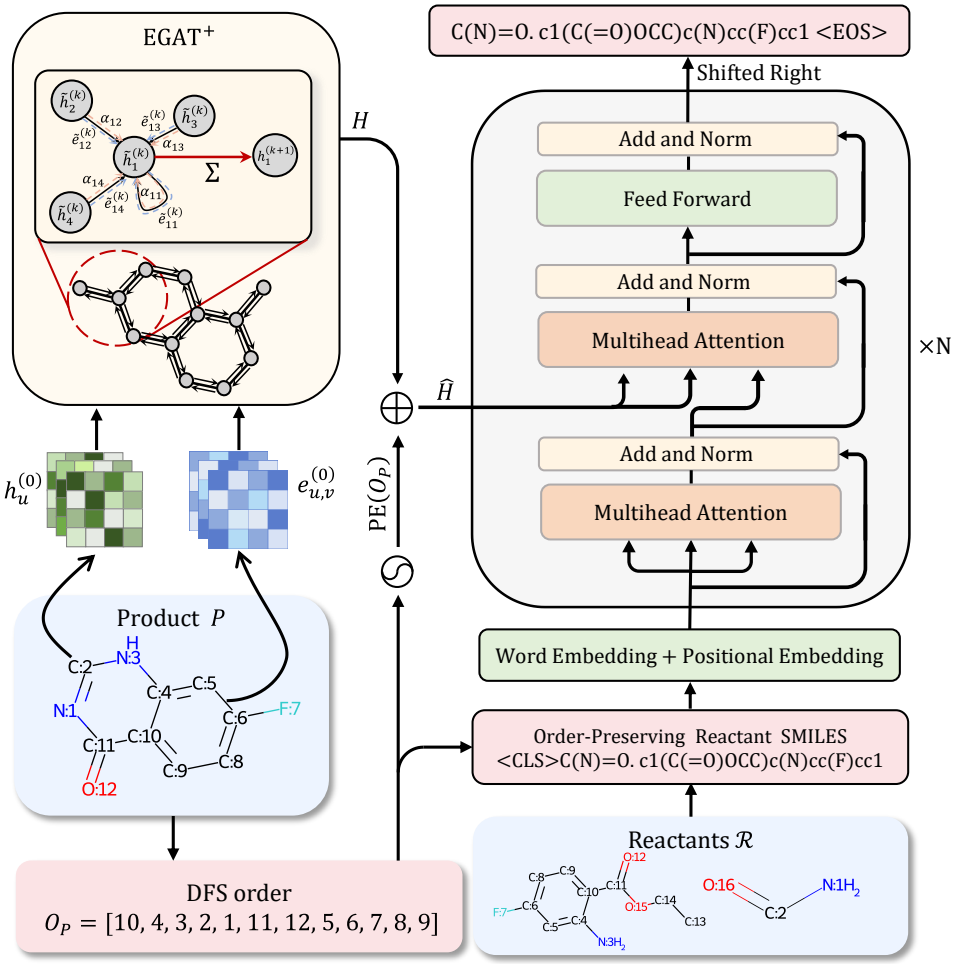


Fig. 1 Overview of UAlign: Given a product molecule graph P and one of its DFS order O_P, the graph is first fed into the graph neural network called EGAT$^+$ to obtain node features H. Then the positional encoding is added to H according to the given DFS order O_P to generate the order-aware node features $\hat{H}$. Finally the decoder takes $\hat{H}$ as input and generate the SMILES of reactants auto-regressively.

written as

$$
\begin{aligned}
\tilde{e}_{u,v}^{(k)} &= \mathrm{FFN}_e^{(k)}(e_{u,v}^{(k)}), \\
\tilde{h}_u^{(k)} &= \mathrm{FFN}_n^{(k)}(h_u^{(k)}), \\
c_{u,v} &= \mathbf{a}^T[\tilde{h}_u^{(k)} \| \tilde{h}_v^{(k)} \| \tilde{e}_{u,v}^{(k)}], \\
\alpha_{u,v} &= \frac{\exp(\mathrm{LeakyReLU}(c_{u,v}))}{\sum_{v' \in \mathcal{N}(u) \cup \{u\}} \exp(\mathrm{LeakyReLU}(c_{u,v'}))}, \\
h_u^{(k+1)} &= \sum_{v \in \mathcal{N}(u) \cup \{u\}} \alpha_{u,v} \left(\tilde{h}_u^{(k)} + \tilde{e}_{u,v}^{(k)} \right), \\
e_{u,v}^{(k+1)} &= \mathrm{FFN}_m^{(k)}([h_u^{(k+1)} \| h_v^{(k+1)} \| e_{u,v}^{(k)}]), u \neq v,
\end{aligned}
\tag{1}
$$

where $\text{FFN}_m^{(k)}$, $\text{FFN}_e^{(k)}$ and $\text{FFN}_n^{(k)}$ are three different feed forward networks, $\mathbf{a}$ is a learnable parameter, $\mathcal{N}(u)$ denotes the one-hop neighbors of node u and $\|$ denotes the concatenation operation. Since there are no chemical bonds with the same beginning and ending atoms, the $e_{u,u}^{(k)}$ is also set as a learnable parameter shared among all atoms. The residual connection and layer normalization [2] are applied to prevent over-smoothing while enlarging the receptive field of the model [36].

The initial node features $h_u^{(0)}$ and edge features $e_{u,v}^{(0)}$ are determined via several chemical property descriptors, whose details are shown in Supplementary Sec. 6. After K iterations of message passing, we can obtain the encoded features $h_u^{(K)}$ of all atoms and make up the output $H \in \mathbb{R}^{V_P \times d}$ of the encoder, where d denotes the embedding size.

2.3 SMILES Alignment

For single-step retrosynthesis prediction, a significant proportion of structures are shared between product molecules and reactant molecules [34, 43]. However, SMILES-based methods often have to generate the reactant SMILES from scratch, even if most of the structures of reactants are the same as those of the products. This results in the underutilization of input information and becomes the bottleneck of template-free retrosynthesis prediction methods. There are methods [25, 34] addressing this issue through supervised SMILES alignment, which involves adding supervised information to establish the correspondence between input and output tokens through cross-attention over the input and predicted tokens. This supervised training approach not only requires complex data annotation algorithms but also limits the diversity of the model's attention map, thereby further affecting the model's performance. To address the above-mentioned issues, we propose the unsupervised SMILES alignment method as follows.

Assuming we can identify the location of each product atom in the reactants' SMILES and provide it to the model, a natural correspondence can be established between the input and output atoms. However, during the inference process, revealing this information would lead to label leakage, which is not permitted. Therefore, we propose the following modification: when providing an order of product atoms, we expect the model to generate atom tokens in the reactants' SMILES in this given order as closely as possible. By doing so, we can establish a correspondence between the product atoms and the reactants' SMILES tokens using unsupervised methods without leaking any labels. We refer to this type of reactants' SMILES, which aims to preserve the given order of atom tokens as much as possible, as order-preserving reactant SMILES. Note that SMILES represents atoms in a molecule according to a certain DFS order, the provided order should also be a DFS order for the product molecule.

The generation of order-preserving reactant SMILES will be introduced as follows. Given the product molecule $P = (V_P, E_P)$ with a DFS order $O_P \in \mathcal{D}(V_P)$ and the corresponding set of reactant molecules $\mathcal{R} = \{R_1, R_2, \dots R_l\}$, for each reactant $R = (V_R, E_R) \in \mathcal{R}$, we can find a depth-first order $O_R \in \mathcal{D}(V_R)$ that has a nearly consistent atomic appearance sequence with O_P as the product and reactants have similar structures. For convenience, we name such a order as O_P-corresponding order

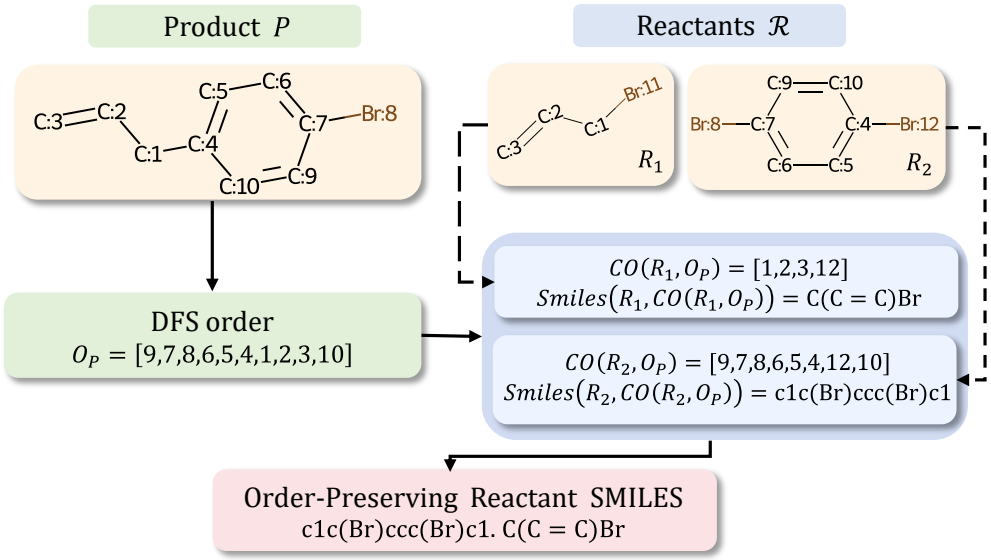


Fig. 2 An example of the process to generate order-preserving reactants SMILES. The atom mapping numbers shown on the figure are included only for clearer explanation and will be removed in our implementation to prevent any label leakage.

of R and denote it as $CO(R, O_P)$. Mathematically, it's defined as

$$CO(R, O_P) = \arg \min_{o \in \mathcal{D}(R)} \sum_{i \in O_P \cap o} \sum_{j \in O_P \cap o} inv(i, j, O_P, o), \tag{2}$$

where the value of $inv(i, j, O_P, o)$ equals 1 if and only if $rank(i, O_P) < rank(j, O_P)$ and $rank(i, o) > rank(j, o)$, and equals 0 otherwise. We sort the reactants $\mathcal{R}$ according to $rank(root(R, CO(R, O_P)), O_P)$ in ascending order. Then we generate SMILES for each reactant molecule using its O_P-corresponding order and join them together using "." to obtain order-preserving reactant SMILES.

For further discussion, we denote the order-preserving reactant SMILES given the reactant molecules $\mathcal{R}$ and a DFS order O of product as $OPSmiles(\mathcal{R}, O)$. An example of the process to generate order-preserving reactants SMILES is shown in Fig. 2. The detailed implementations are presented in Supplementary Sec. 5.1.

2.4 Decoder

The decoder takes the node features $H \in \mathbb{R}^{V_P \times d}$ that are generated from the encoder, as well as the given DFS order O_P for the product molecule graph as input. We use the vanilla Transformer decoder [31] as our decoder. As mentioned in Sec. 2.3, the order information of product atoms are required for SMILES alignment. However, the Transformer decoder is permutation-invariant to memory [17, 29], meaning it is not sensitive to the order of the features from encoder. This implies that directly performing

cross-attention over H may not effectively capture the relationship between product atoms and reactant SMILES tokens. To address this problem, we introduce position encoding to the node features based on the rank of each atom in the given DFS order O_P to generate order-aware node features $\hat{H}$. Then given an input embedding sequence $Z \in \mathbb{R}^{m \times d}$, the Transformer decoder layer utilizes the order-aware node features as keys and values in all the cross-attention layers. This process ultimately generates the decoded embeddings $\hat{Z} \in \mathbb{R}^{m \times d}$. These embeddings are then fed into feed-forward layers $\mathrm{FFN}_1 : \mathbb{R}^d \to \mathbb{R}^T$ to predict the tokens $\hat{T}$ that should be generated. In summary, the decoder can be mathematically expressed as

$$\begin{aligned} \hat{H} &= H + \mathrm{PE}(O_P), \\ \hat{Z} &= \mathrm{TransformerDecoder}(Z, \hat{H}), \\ \hat{T} &= \mathrm{FFN}_1(\hat{Z}). \end{aligned} \tag{3}$$

2.5 Two stage training

There is a significant distribution shift between graphs and SMILES representations. Moreover, our model is specifically designed to generate non-canonical SMILES, which may contain more complex patterns compared to canonical SMILES. To achieve this goal, we propose a two-stage training strategy in this paper. The first stage aims to align the distributions between two distinct modalities: SMILES and molecular graphs, while enabling the model to learn the patterns of non-canonical SMILES. Given a molecule graph M and one of its possible DFS orders O_M, the training task is to translate the graph into the corresponding SMILES representation based on the given order O_M. In detail, this is reached by training the model to generate $Smiles(M, O_M)$ given molecule M and DFS order O_M.

Once the first stage training converges, we proceed to the second stage, which focuses on retrosynthesis prediction. In this stage, the model is trained using the order-preserving reactant SMILES as targets. Given a product molecule graph P, a possible DFS order O_P, and a set of reactants $\mathcal{R}$, the model is expected to generate $OPSmiles(\mathcal{R}, O_P)$.

2.6 Data Augmentation

Different from those Transformer-based methods [25, 28, 34] taking SMILES as input and canonical SMILES as target, our method takes a graph as input and is trained with non-canonical SMILES. That means the previous SMILES augmentation tricks are not suitable for us. Similar to [34], we choose to augment the training data on-the-fly.

For the first stage, at each iteration, for each molecule $M = (V_M, E_M)$, we have a 50% chance of using a random DFS order O_M as the input for the model, and using the corresponding $Smiles(M, O_M)$ as the training target. For the other 50%, we randomly select another molecule $M' = (V_{M'}, E_{M'})$ from the dataset to form a new molecular graph $\tilde{M} = (V_M \cup V_{M'}, E_M \cup E_{M'})$, and find the DFS order $O_{\tilde{M}}$ that can generate canonical SMILES for $\tilde{M}$. $\tilde{M}$ and $O_{\tilde{M}}$ are then fed into the model and the target is set as the canonical SMILES of $\tilde{M}$. Such an augmentation method enables the model

to output the atom tokens according to the given DFS order and be aware of different components within a graph.

For the second stage, at each iteration, for each product molecule $P = (V_P, E_P)$, we have a 50% probability of using a random DFS order as input, and for the remaining part, we use the DFS order capable of producing canonical SMILES for product as input. The target used for training is the order-preserving reactant SMILES generated based on the input DFS order. This data augmentation method allows the model to focus more on the DFS order for canonical product SMILES while also noticing the correspondence between product atoms and the output SMILES tokens.

2.7 Loss

Both stage of training can considered as a kind of translation between graphs and SMILES, thus we use the loss widely used for auto-regressive language generation models for training. Denote the training target as $T = \{t_1, t_2, \ldots, t_n\}$ and the output of the model $\hat{T} = \{\hat{t}_1, \hat{t}_2, \ldots, \hat{t}_n\}$, the loss can be written as

$$\mathcal{L} = \sum_{i=1}^{n} l_{cls}(\hat{t}_i, t_i), \tag{4}$$

where $l_{cls}(\cdot)$ is the classification loss.

3 Results and Discussion

In this section, we conduct extensive experiments to make a comprehensive evaluation of our proposed UAlign.

3.1 Evaluation Protocol

Benchmark Datasets. We adopt three datasets for evaluation: (1) **USPTO-50K** consists of 50,016 atom-mapped reactions grouped into 10 different classes; (2) **USPTO-FULL** comprises 1,013,118 atom-mapped reactions without any reaction class information. (3) **USPTO-MIT** consists of 479,035 atom-mapped reactions without any reaction class information. To ensure a fair comparison, we adopt the same training/validation/test splits as those in a previous study [9] for USPTO-50K and USPTO-FULL datasets. The training/validation/test splits is aligned with the previous study [14]. The detailed data processing procedure and the statistical information of the processed dataset are presented in Supplementary Sec. 2.

Metrics. We utilize the following three evaluation metrics for evaluation: **top-k accuracy**, **top-k SMILES validity**, and **top-k round-trip accuracy**. The detailed definitions for three metrics are provided in Supplementary Sec. 3.

3.2 Performance Comparison

Top-k Accuracy. We compare our model with existing single-step retrosynthesis prediction in terms of top-k accuracy on all the datasets. The results are summarized in

Table 1 Top-k accuracy for retrosynthesis prediction on USPTO-50K. * indicates the model with SMILES augmentation. For comparison purpose, the Aug. Transformer is evaluated without the test augmentation. Best performance of each model type is in **bold**.

Model	Top-k accuracy (%)							
	Reaction class known				Reaction class unknown			
	1	3	5	10	1	3	5	10
Template-Based								
RetroSim [7]	52.9	73.8	81.2	88.1	37.3	54.7	63.3	74.1
NeuralSym [24]	55.3	76.0	81.4	85.1	44.4	65.3	72.4	78.9
GLN [9]	64.2	79.1	85.2	90.0	52.5	69.0	75.6	83.7
LocalRetro [5]	63.9	86.8	92.4	96.3	53.4	77.5	85.9	92.4
RetroKNN [37]	**66.7**	**88.2**	**93.6**	**96.6**	**57.2**	**78.9**	**86.4**	**92.7**
Semi-Template-Based								
RetroXpert* [38]	62.1	75.8	78.5	80.9	50.4	61.1	62.3	63.4
G2G [26]	61.0	81.3	86.0	88.7	48.9	67.6	72.5	75.5
GraphRetro [27]	63.9	81.5	**85.2**	**88.1**	**53.7**	68.3	72.2	75.5
RetroPrime* [35]	**64.8**	**81.6**	85.0	86.9	51.4	**70.8**	**74.0**	**76.1**
Template-Free								
Transformer [31]	57.1	71.5	75.0	77.7	42.4	58.6	63.8	67.7
SCROP [42]	59.0	74.8	78.1	81.1	43.7	60.0	65.2	68.7
Liu's Seq2seq [19]	-	-	-	-	37.4	52.4	57.0	61.7
Tied Transformer [15]	-	-	-	-	47.1	67.1	73.1	76.3
Aug. Transformer* [28]	-	-	-	-	48.3	-	73.4	77.4
MEGAN [21]	60.7	82.0	87.5	91.6	48.1	70.7	78.4	86.1
GTA* [25]	-	-	-	-	51.1	67.6	74.8	81.6
Graph2SMILES [29]	-	-	-	-	52.9	66.5	70.0	72.9
Retroformer* [34]	64.0	82.5	86.7	90.2	53.2	71.1	76.6	82.1
RetroBridge [13]	-	-	-	-	50.8	74.1	80.6	85.6
Ours*	**66.2**	**86.9**	**91.9**	**95.1**	**53.6**	**77.6**	**84.6**	**90.3**

Table 2 Top-k accuracy for retrosynthesis prediction on USPTO-MIT. Best performance of each model type is in **bold**.

Model Type	Model	Top-k accuracy (%)			
		1	3	5	10
template-based	NeuralSym [24]	47.8	67.6	74.1	80.2
	LocalRetro [5]	**54.1**	**73.7**	**79.4**	**84.4**
template-free	Liu's Seq2seq [19]	46.9	61.6	66.3	70.8
	AutoSynRoute [18]	54.1	71.8	76.9	81.8
	RetroTRAE [30]	58.3	-	-	-
	Ours	**59.9**	**76.9**	**82.0**	**86.4**

Table 1, Table 2 and Table 3. On the USPTO-50K dataset, our model achieves a top-3 accuracy of 77.6%, top-5 accuracy of 84.6% and top-10 accuracy of 90.3% under the reaction class unknown setting, surpassing the SOTA template-free method by 3.5%, 4.0% and 4.7% respectively. And with reaction class given on USPTO-50K dataset, our model achieves the top-1 accuracy of 66.2%, top-5 accuracy of 91.9% and top-10 accuracy of 95.1%, which exceeds the SOTA template-free method by 2.2%, 4.4% and 4.9% respectively. Moreover, our model outperforms all the semi-template-based methods with a noticeable margin. It's also encouraging to see that our method, as a

Table 3 Top-k accuracy for retrosynthesis prediction on USPTO-FULL. * indicates the model with SMILES augmentation. Best performance of each model type is in **bold**.

Model Type	Model	Top-k accuracy (%)			
		1	3	5	10
template-based	RetroSim [7]	32.8	-	-	56.1
	NeuralSym [24]	35.8	-	-	60.8
	GLN [9]	**39.3**	-	-	**63.7**
	LocalRetro [5]	39.1	**53.3**	**58.4**	63.7
semi-template-based	RetroPrime* [35]	**44.1**	**59.1**	**62.8**	**68.5**
template-free	MEGAN [21]	33.6	-	-	63.9
	Aug. Transformer* [28]	46.2	-	-	73.3
	Graph2SMILES [29]	45.7	-	-	63.4
	GTA* [25]	46.6	-	-	70.4
	Ours*	**50.4**	**66.1**	**71.3**	**76.2**

Table 4 Top-k SMILES validity for retrosynthesis prediction on USPTO-50K with reaction class unknown.

Model	Top-k validity (%)			
	1	3	5	10
Transformer [31]	97.2	87.9	82.4	73.1
Graph2SMILES [29]	99.4	90.9	84.9	74.9
RetroPrime [35]	98.9	98.2	97.1	92.5
Retroformer [34]	99.2	98.5	97.4	96.7
Ours	**99.8**	**99.7**	**99.3**	**98.2**

Table 5 Top-k round-trip accuracy for retrosynthesis prediction on USPTO-50K with reaction class unknown.

Model	Top-k round-trip acc. (%)			
	1	3	5	10
GraphRetro [27]	80.5	73.3	68.3	59.3
Transformer [31]	71.9	54.7	46.2	35.6
Graph2SMILES [29]	76.7	56.0	46.4	34.9
RetroPrime [35]	79.6	59.6	50.3	40.4
Retroformer [34]	78.9	72.0	67.1	57.2
Ours	**80.9**	**74.0**	**69.0**	**60.2**

template-free method, achieves competitive or even superior performance against the powerful template-based methods such as LocalRetro under both settings of USPTO-50K dataset. On UPSTO-MIT dataset, our model achieves the top-1 accuracy of 59.9% and top-10 accuracy of 86.4%, which even outperforms the existing template-based SOTA method LocalRetro significantly. Additionally, our models achieved a top-1 accuracy of 50.4% on the USPTO-FULL dataset, which exceed that of the current SOTA model GTA by 3.8%. These findings sufficiently demonstrate the effectiveness of our method. The contribution of each proposed module will be further validated in Sec. 3.3.

It is noteworthy that while template-based approaches have achieved remarkable performance on the USPTO-50K dataset, their reliance on external template libraries has emerged as a constraint as datasets grow in scale and complexity. This dependency leads to a substantial degradation in model performance. In contrast, template-free methods have demonstrated superior versatility and adaptability, qualities that render them especially appropriate for managing large-scale and intricate datasets.

Top-k SMILES Validity. We use vanilla Transformer, RetroPrime, Retroformer and Graph2SMILES as robust baselines to compare the validity of SMILES in our study. SMILES generation models for retrosynthesis tasks often encounter challenges with maintaining SMILES validity. We do not take the methods based on templates or molecule editing as baseline here because the validity of generated SMILES can be guaranteed by the templates or chemical toolkits. Unlike graph-based models, SMILES-based methods need to ensure that the generated content adheres to the parsing rules of SMILES, without leveraging chemical tools such as RDKit. Consequently, SMILES-based approaches are more susceptible to generating invalid SMILES compared to graph-based models. As shown in Table 4, our model demonstrates superior top-1 and top-5 molecule validity compared to other models, even without employing canonical SMILES as our training objective. This improvement could be attributed to the proposed two-stage training strategy and data augmentation, which assist the model in capturing various SMILES patterns effectively.

Top-k Round-Trip Accuracy. To assess the accuracy of our predicted synthesis plans, we utilize the Molecule Transformer [23] as the benchmark reaction prediction model and calculate the top-k round-trip accuracy. We take RetroPrime, Retroformer and Graph2SMILES as our strongv SMILES-based baselines. We also use take graph-based method GraphRetro into comparison. The results are presented in Table 5. The results clearly indicate that our model outperforms all SMILES-based baselines by a considerable extent and even exceeds the well-established graph-based method, GraphRetro. This underscores the efficacy of our unsupervised SMILES alignment mechanism, which enables the model to efficiently leverage substructures from product molecules to construct reactants. This mechanism allows the model to focus more intently on learning reaction mechanisms, thereby yielding more plausible predictive outcomes. In summary, our model has exhibited a robust capacity for generating coherent and efficacious synthesis pathways, specifically tailored for advanced downstream applications such as multi-step retrosynthesis planning.

3.3 Ablation Study

We investigate the effects of different components in our proposed pipelines. The result is summarized in Table 6.

Two Stage Training. We eliminate the initial training phase and directly train the model for the retrosynthesis prediction task. As indicated in Table 6, the two-stage training strategy has consistently led to enhancements in all evaluated metrics. This observation implies that the two-stage training strategy effectively enables the model to adeptly learn the intricacies of molecular SMILES representations, thereby yielding higher quality and more plausible retrosynthetic analysis outcomes.

Table 6 Effects of different modules on retrosynthesis performance in reaction class unknown setting of USPTO-50K dataset. Best performance is in **bold**.

Method	**Top-k acc. (%)**				**Top-k round-trip acc. (%)**				**Top-k validity (%)**			
	1	3	5	10	1	3	5	10	1	3	5	10
UAlign (Full Version)	**53.6**	**77.6**	**84.6**	**90.3**	**80.9**	**74.0**	**69.0**	**60.2**	**99.8**	**99.7**	**99.3**	**98.2**
– Two Stage Training	53.1	75.6	82.8	88.8	80.4	72.3	66.9	57.6	99.2	97.7	96.1	92.2
– Data Augmentation	48.5	71.8	79.2	85.9	78.2	70.5	65.1	56.0	99.1	98.4	97.7	95.6
– SMILES Alignment	45.5	65.2	71.4	77.5	70.3	57.9	51.0	40.1	97.8	95.3	93.6	89.2
Transformer [31]	42.4	58.6	63.8	67.7	71.9	54.7	46.2	35.6	97.2	87.9	82.4	73.1

Data Augmentation. We remove the data augmentation during the second training stage, which means training solely using the DFS order that can generate canonical SMILES. Table 6 demonstrates a significant decline in model performance across all metrics. This clearly demonstrates that our data augmentation significantly improves the model's performance.

SMILES Alignment. In the training process, we remove all operations related to SMILES alignment. This includes the removal of the position encoding in Eq. 3, where the features H directly served as the input memory for the Transformer decoder. Since we eliminate the input related to the DFS order, the model was no longer trained using order-preserving reactants SMILES as the target but instead switched to canonical SMILES for product. Additionally, in this set of experiments, we remove the first training stage, which aligns the graph and SMILES modalities as the model architecture changes. The results are reported in Table 6, and they show a significant decline in performance compared to our full version, indicating that the proposed SMILES alignment algorithm is crucial for achieving excellent performance.

It is worth noting that even without data augmentation, two-stage training and SMILES alignment, our model still outperforms the vanilla Transformer by a large margin in terms of all metrics reported in the last line of Table 6. This indicates that graph-based molecular representation learning still has advantages over SMILES-based approaches, and our proposed EGAT$^+$ can extract effective molecular representations for downstream usage.

3.4 Case Study (Visualization of cross-attention mechanism in transformer with UAling)

We randomly select a case from the dataset and showcase the cross-attention map in Fig. 3. The cross-attention map indicates the correlation between reactant tokens and nodes in the input product graph. This map is obtained by averaging the attention coefficient from each attention head. From the figure, it is evident that the predicted tokens successfully locate their corresponding atoms in the product, which contributes to the accurate prediction. The SMILES alignment can also be observed to assist the model in correctly identifying the reaction center. In accordance with the figure, the bond between atom `C:11` and `N:9` breaks during the transformation into reactants. Our model effectively notices this occurrence and focus the attention of token $\hat{t}_7$ on the reaction centers `C:11` and `N:9`. This strategic focusing successfully guides the completion of the reactants, ensuring that the leaving group is correctly attached to the appropriate atoms. Additionally, we note that the attention coefficient at token $\hat{t}_{14}$ is concentrated on atoms `C:1` and `N:9`, which are the first atoms of each reactant molecule according to the given DFS order. This further indicates that our model is able to correctly identify the sites where the reaction occurs and accurately cleave the chemical bonds. Moreover, the attention of newly generated structures (i.e., tokens $\hat{t}_7$ to $\hat{t}_{13}$) is directed towards atoms `C:1`, `C:11`, and `O:2`, which correspond to the specific synthon they will attach to. This demonstrates that our model is able to generate appropriate functional groups based on the molecular structure information to form the reactants. All the aforementioned results illustrate that our proposed SMILES

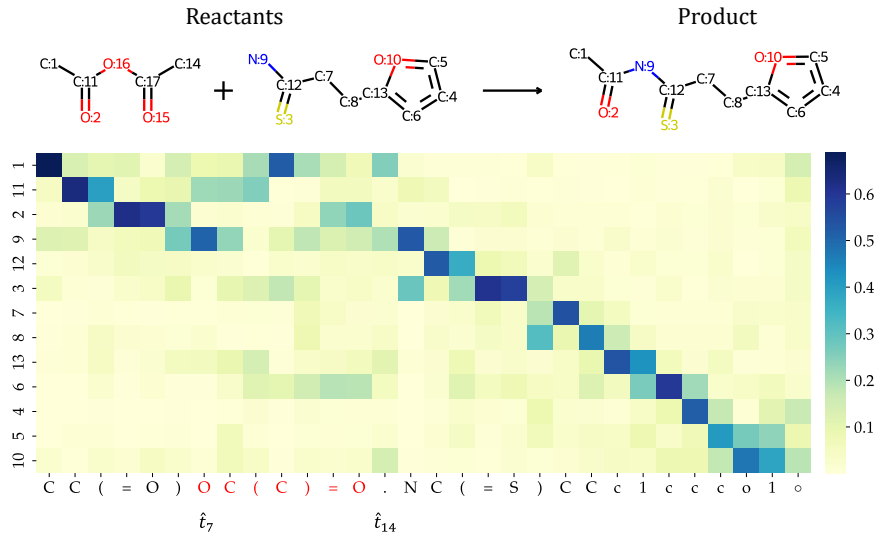


Fig. 3 Visualization of cross-attention over order-aware node features and the predicted tokens. The number on the y-axis is the map number of atoms in the product. The reactants atoms that not appear in product is colored red in the x-axis. ∘ represents the end token.

alignment method assists the model in comprehending molecular structural information and helps it to focus on learning chemical rules.

To further investigate the impact of the proposed SMILES alignment mechanism on model training, we visualize the cross-attention coefficients of different Transformer decoder layers. The visualization is provided in Supplementary Fig. 1 of Supplementary Information. From Supplementary Fig. 1, we can observe significant variations in the cross-attention across different layers. Additionally, the establishment of correspondence does not occur exclusively at certain layers, such as the first or last layer. This suggests that directly imposing supervised signals on the cross attention coefficient [25, 34] for SMILES alignment is not a wise approach, whether applied to all layers or only the last layer. It also verifies our statement in Sec. 2.3 that using unsupervised methods for SMILES alignment does not affect the diversity of attention maps and further has a negative impact on model training and performance. This is why our unsupervised SMILES alignment mechanism achieves better results than supervised SMILES alignment.

3.5 Case Study (Multi-step Retrosynthetic Pathway Planning)

To explore the suitability of our model for multi-step retrosynthetic pathway planning, we select three distinct molecules as targets for synthetic route design, and the synthesis routes are obtained through iterative calls to our UAlign model, which is trained with the USPTO-FULL dataset. The predicted pathways are summarized in Fig. 4.

The first case study involves Mitapivat, a compound approved for the treatment of hereditary hemolytic anemias in February 2022 [1]. Our model successfully predicted

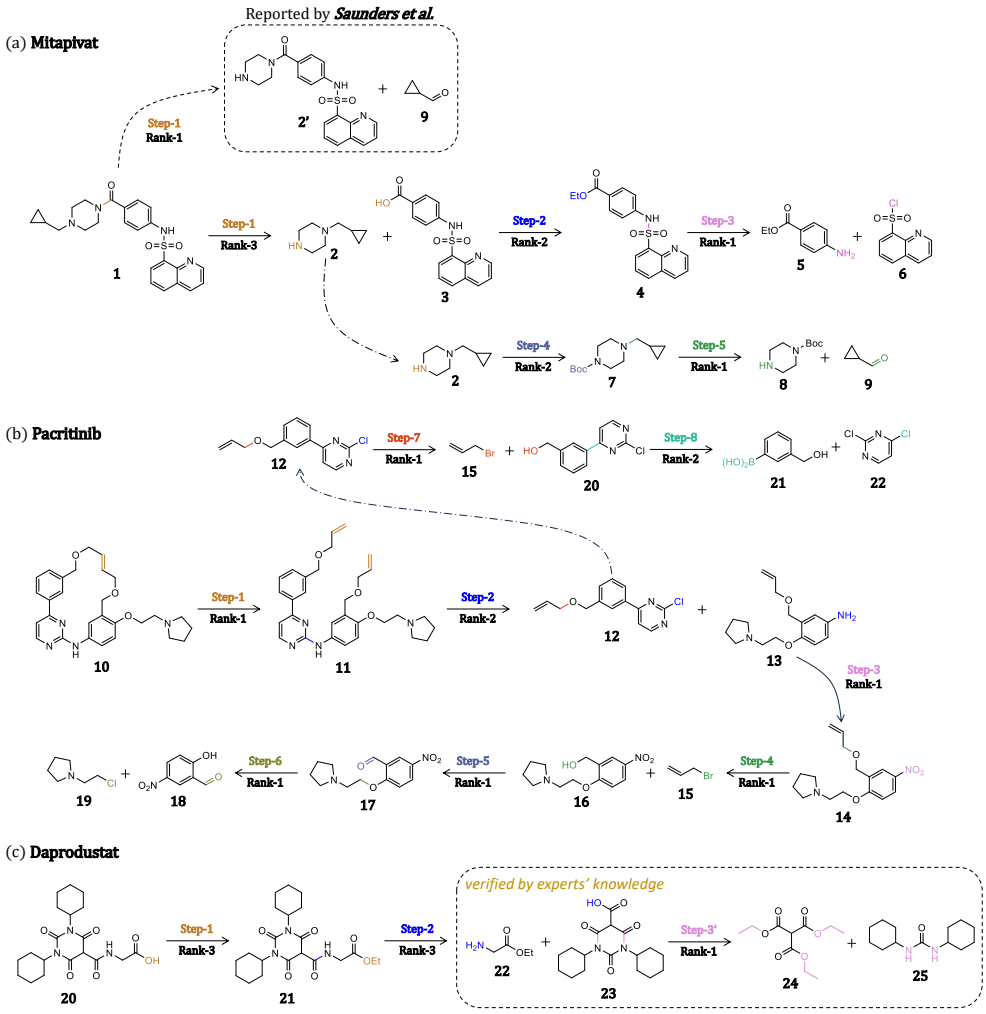


Fig. 4 Multistep retrosynthesis predictions by our method. (a) Mitapivat (b) Pacritinib citrate (c) Daprodust. The reaction centers and leaving groups are highlighted in different colors. The pathway pf molecule (a) and (b) come from literature, while the last one is verified by chemical experts.

the five-step synthetic route reported in [3], with each step consistently ranked within the top-2 predictions. The first step entails an amide coupling reaction, which our model placed at rank 2, yielding the reactants 1-(cyclopropylmethyl)piperazine (compound **2**) and 4-(quinoline-8-sulfonamido)benzoic acid (compound **3**). Notably, at the initial step, our model also proposed an alternative synthesis method utilizing the Borch Reductive Amination, which was ranked at the first and is consistent with the synthetic route delineated by Saunders et al. Subsequently, for the synthesis of 4-(quinoline-8-sulfonamido)benzoic acid, the model precisely executed a functional group protection strategy during the second step and accurately anticipated the subsequent formation

of the sulfonamide, effectively deconstructing the target molecule into readily available precursors. For the synthesis of 1-(cyclopropylmethyl)piperazine, the model strategically protected the amine functional group with a tert-butyloxycarbonyl moiety at the outset and, in the ultimate step, prognosticated the N-alkylation reaction with a top-ranking accuracy. This example illustrates our model's capability to uncover diverse reaction centers in molecular retrosynthetic design and to generate plausible reactant combinations based on these insights.

The second case under scrutiny is Pacritinib, an orally bioavailable and isoform selective JAK-2 inhibitor for the treatment of patients with myelofibrosis, which received FDA approval on February 28, 2022 [41]. As shown in Figure 4(b), our model successfully delineates a eight-step synthesis, as described in the literature [4], tracing the synthetic pathway from commercially available 5-nitrosalicylaldehyde and 2,6-dichloropyrimidine to the final product. The initial step of the reverse synthesis is olefin metathesis, ranking the first in order of likelihood, followed by another rank-2 aromatic substitution of 4-(3-((allyloxy)methyl)phenyl)-2-chloropyrimidine (compound **12**) and 3-((allyloxy)methyl)-4-(2-(pyrrolidin-1-yl)ethoxy)aniline (compound **13**). Subsequently, synthesis of 4-(3-((allyloxy)methyl)phenyl)-2-chloropyrimidine was correctly identified via continuous allyl substitution and Suzuki cross-coupling reaction as the top and the second choices. The reverse synthesis of 3-((allyloxy)methyl)-4-(2-(pyrrolidin-1-yl)ethoxy)aniline was reduction of the nitro group, followed by another allyl substitution. In the final step, the model's highest probability prediction was reduction of the aldehyde group, followed by a nucleophilic substitution. Despite the synthesis route involving a considerable number of steps and encompassing a variety of reaction types, our model successfully and accurately predicted each step within the top-2 choices. This accomplishment signifies the robustness and efficacy of our model in the context of retrosynthetic analysis.

The final case is Daprodustat, the first oral hypoxy-inducing factor prolyl hydroxylase inhibitor (HIF-PHI) for the treatment of renal anemia caused by chronic kidney disease (CKD) [11]. This novel compound received approval for market release from the FDA on the 1st of February, 2023 [41]. Our model predicted the three-step synthetic route. The first step reports the hydrolysis of ester at rank 3, which is aligned with the route provided by Duffy et al. Although next two steps provided by our method do not exist in the literature, there are all explainable. The synthesis of ethyl (1,3-dicyclohexyl-2,4,6-trioxohexahydropyrimidine-5-carbonyl)glycinate (compound **21**) was identified via dehydration condensation of 1,3-dicyclohexyl-2,4,6-trioxohexahydropyrimidine-5-carboxylic acid (compound **23**) and ethyl glycinate (compound **22**) as the top choice, which avoided using toxic ethyl isocyanatoacetate reported in literature. In the final step, the model's highest probability prediction was amidation of ester, resulting in cost-effective and readily accessible starting materials. This case demonstrates the robust extrapolative capacity of our model, highlighting its potential to generate synthetic routes that surpass those documented in the literature.

4 Limitations

This work does not integrate much domain knowledge related to chemical reaction mechanisms in its design, which to some extent, compromises its interpretability. Similarly to most template-free methods, our work also faces challenges in generating diverse results. These aspects will be left for exploration in future works.

5 Conclusion

We present UAlign, a novel graph-to-sequence pipeline that achieves state-of-the-art performance in the field of template-free methods. Our approach outperforms existing template-free and semi-template-based methods, while achieving comparable results to template-based methods. By utilizing a specially-designed graph neural network as the encoder, our model effectively leverages chemical and structural information from molecule graphs, resulting in powerful embedding for the decoder. Additionally, Our proposed unsupervised SMILES alignment mechanism facilitates the reuse of shared substructures between reactants and products for generation, allowing the model to prioritize chemical knowledge even without complex data annotations. This significantly enhances the performance of the pipeline. In future work, we plan to further explore multi-step retrosynthesis planning using our UAlign as the single-step retrosynthesis prediction backbone.

Supplementary information. We provide the details about our metrics, data preprocess and more visualization results in another pdf file as Supplementary Information.

Declarations

5.1 Funding

This work was supported by the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), the SJTU AI for Science platform, and the Fundamental Research Funds for the Central Universities.

5.2 Competing interests

The authors declare no competing interests.

5.3 Data availability

The Data of USPTO-FULL and USPTO-50K can be found in https://github.com/Hanjun-Dai/GLN. The Data of USPTO-MIT can be found in https://github.com/wengong-jin/nips17-rexgen/tree/master/USPTO. We also provide the raw data and our processed version in https://github.com/zengkaipeng/UAlign.

5.4 Code availablity

All the codes and checkpoints can be found in https://github.com/zengkaipeng/UAlign.

5.5 Author contribution

K.Z. was responsible for the code implementation, algorithm design, and the majority of the manuscript writing. X.Z. conducted the Ablation Study. Y.Z. and B.Y. managed multi-step Retrosynthetic Pathway Planning. F.N. handled all visualizations. The remaining authors contributed to the polishing of the article. Y.X., Y.J. and X.Y. supervised the research.

5.6 Ethics approval and consent to participate

Not applicable

5.7 Materials availability

Not applicable

5.8 Consent for publication

Not applicable

References

[1] Al-Samkari H, van Beers EJ (2021) Mitapivat, a novel pyruvate kinase activator, for the treatment of hereditary hemolytic anemias. Therapeutic Advances in Hematology 12:20406207211066070. https://doi.org/10.1177/20406207211066070, URL https://doi.org/10.1177/20406207211066070, pMID: 34987744, https://doi.org/10.1177/20406207211066070

[2] Ba JL, Kiros JR, Hinton GE (2016) Layer normalization

[3] Benedetto Tiz D, Bagnoli L, Rosati O, et al (2022) Fda-approved small molecules in 2022: Clinical uses and their synthesis. Pharmaceutics 14(11):2538

[4] Chang H, Yajun G, Jiajuan P, et al (2015) Synthesis of pacritinib hydrochloride. Chinese Journal of Pharmaceuticals 46(12). https://doi.org/10.16522/j.cnki.cjph.2015.12.001

[5] Chen S, Jung Y (2021) Deep retrosynthetic reaction prediction using local reactivity and global attention. JACS Au 1(10):1612–1620

[6] Chen Z, Ayinde OR, Fuchs JR, et al (2023) G 2 retro as a two-step graph generative models for retrosynthesis prediction. Communications Chemistry 6. https://doi.org/10.1038/s42004-023-00897-3

[7] Coley CW, Rogers L, Green WH, et al (2017) Computer-assisted retrosynthesis based on molecular similarity. ACS central science 3(12):1237–1245

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

[9] Dai H, Li C, Coley C, et al (2019) Retrosynthesis prediction with conditional graph logic network. Advances in Neural Information Processing Systems 32

[10] Duffy K, Fitch D, Jin J, et al (2007) Preparation of n-substituted pyrimidine-trione amino acid derivatives as prolyl hydroxylase inhibitors. WO2007150011A2

[11] Hara K, Takahashi N, Wakamatsu A, et al (2015) Pharmacokinetics, pharmacodynamics and safety of single, oral doses of gsk1278863, a novel hif-prolyl hydroxylase inhibitor, in healthy japanese and caucasian subjects. Drug Metabolism and Pharmacokinetics 30(6):410–418. https://doi.org/https://doi.org/10.1016/j.dmpk.2015.08.004, URL https://www.sciencedirect.com/science/article/pii/S1347436715000518

[12] Hu W, Liu B, Gomes J, et al (2019) Strategies for pre-training graph neural networks. In: International Conference on Learning Representations

[13] Igashov I, Schneuing A, Segler M, et al (2023) Retrobridge: Modeling retrosynthesis with markov bridges. In: The Twelfth International Conference on Learning Representations

[14] Jin W, Coley C, Barzilay R, et al (2017) Predicting organic reaction outcomes with weisfeiler-lehman network. Advances in neural information processing systems 30

[15] Kim E, Lee D, Kwon Y, et al (2021) Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. Journal of Chemical Information and Modeling 61(1):123–133

[16] Landrum G, et al (2013) Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 8:31

[17] Lee J, Lee Y, Kim J, et al (2019) Set transformer: A framework for attention-based permutation-invariant neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol 97. PMLR, pp 3744–3753, URL https://proceedings.mlr.press/v97/lee19d.html

[18] Lin K, Xu Y, Pei J, et al (2020) Automatic retrosynthetic route planning using template-free models. Chemical science 11(12):3355–3364

[19] Liu B, Ramsundar B, Kawthekar P, et al (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS central science 3(10):1103–1113

[20] Rong Y, Bian Y, Xu T, et al (2020) Self-supervised graph transformer on large-scale molecular data. In: Larochelle H, Ranzato M, Hadsell R, et al (eds) Advances in Neural Information Processing Systems, vol 33. Curran Associates, Inc., pp 12559–12571, URL https://proceedings.neurips.cc/paper_files/paper/2020/file/94aef38441efa3380a3bed3faf1f9d5d-Paper.pdf

[21] Sacha M, Błaz M, Byrski P, et al (2021) Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. Journal of Chemical Information and Modeling 61(7):3273–3284

[22] Saunders JO, Salituro FG, Yan S (2010) Preparation of aroylpiperazines and related compounds as pyruvate kinase m2 modulators useful in treatment of cancer. US2010331307A1, 30 Dec 2010

[23] Schwaller P, Laino T, Gaudin T, et al (2019) Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS central science 5(9):1572–1583

[24] Segler MH, Waller MP (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry–A European Journal 23(25):5966–5971

[25] Seo SW, Song YY, Yang JY, et al (2021) Gta: Graph truncated attention for retrosynthesis. Proceedings of the AAAI Conference on Artificial Intelligence 35(1):531–539. https://doi.org/10.1609/aaai.v35i1.16131, URL https://ojs.aaai.org/index.php/AAAI/article/view/16131

[26] Shi C, Xu M, Guo H, et al (2020) A graph to graphs framework for retrosynthesis prediction. In: International conference on machine learning, PMLR, pp 8818–8827

[27] Somnath VR, Bunne C, Coley C, et al (2021) Learning graph models for retrosynthesis prediction. Advances in Neural Information Processing Systems 34:9405–9415

[28] Tetko IV, Karpov P, Van Deursen R, et al (2020) State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. Nature communications 11(1):5575

[29] Tu Z, Coley CW (2022) Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. Journal of chemical information and modeling 62(15):3503–3513

[30] Ucak UV, Ashyrmamatov I, Ko J, et al (2022) Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. Nature communications 13(1):1186

[31] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. Advances in neural information processing systems 30

[32] Veličković P, Cucurull G, Casanova A, et al (2018) Graph attention networks. In: International Conference on Learning Representations

[33] Vijayakumar AK, Cogswell M, Selvaraju RR, et al (2016) Diverse beam search: Decoding diverse solutions from neural sequence models. arXiv preprint arXiv:161002424

[34] Wan Y, Hsieh CY, Liao B, et al (2022) Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In: International Conference on Machine Learning, PMLR, pp 22475–22490

[35] Wang X, Li Y, Qiu J, et al (2021) Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. Chemical Engineering Journal 420. https://doi.org/10.1016/j.cej.2021.129845

[36] Wu Q, Zhao W, Li Z, et al (2022) Nodeformer: A scalable graph structure learning transformer for node classification. Advances in Neural Information Processing Systems 35:27387–27401

[37] Xie S, Yan R, Guo J, et al (2023) Retrosynthesis prediction with local template retrieval. Proceedings of the AAAI Conference on Artificial Intelligence 37(4):5330–5338. https://doi.org/10.1609/aaai.v37i4.25664, URL https://ojs.aaai.org/index.php/AAAI/article/view/25664

[38] Yan C, Ding Q, Zhao P, et al (2020) Retroxpert: Decompose retrosynthesis prediction like a chemist. Advances in Neural Information Processing Systems 33:11248–11258

[39] Yang N, Zeng K, Wu Q, et al (2023) Molerec: Combinatorial drug recommendation with substructure-aware molecular representation learning. In: Proceedings of the ACM Web Conference 2023, pp 4075–4085

[40] Yao L, Guo W, Wang Z, et al (2024) Node-aligned graph-to-graph: Elevating template-free deep learning approaches in single-step retrosynthesis. JACS Au

[41] Zhang JY, Wang YT, Sun L, et al (2023) Synthesis and clinical application of new drugs approved by fda in 2022. Molecular Biomedicine 4(1):26

[42] Zheng S, Rao J, Zhang Z, et al (2019) Predicting retrosynthetic reactions using self-corrected transformer neural networks. Journal of chemical information and modeling 60(1):47–55

[43] Zhong Z, Song J, Feng Z, et al (2022) Root-aligned smiles: a tight representation for chemical reaction prediction. Chemical Science 13(31):9023–9034

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SIJcheminfoUAlignzengkaipeng.zip