Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Special Track on AI4Tech: AI Enabling Critical Technologies

KGCL: Knowledge-Enhanced Graph Contrastive Learning for Retrosynthesis Prediction Based on Molecular Graph Editing

Fengqin Yang[1] , **Dekui Zhao**[1] , **Haoxuan Qiu**[1] , **Yifei Li**[2] and **Zhiguo Fu**[1] *
[1]School of Information Science and Technology , Northeast Normal University
[2]Department of Chemistry , Northeast Normal University
{yangfq147, zhaodekui, hxqiu, liyf640, fuzg432}@nenu.edu.cn

Abstract

Retrosynthesis, which predicts the reactants of a given target molecule, is an essential task for drug discovery. Retrosynthesis prediction based on molecular graph editing has garnered widespread attention due to excellent interpretability. Existing methods fail to effectively incorporate the chemical knowledge when learning molecular representations. To address this issue, we propose a Knowledge-enhanced Graph Contrastive Learning model (KGCL), which retrieve functional group embeddings from a chemical knowledge graph and integrate them into the atomic embeddings of the product molecule using an attention mechanism. Furthermore, we introduce a graph contrastive learning strategy that generates augmented samples using graph edits to improve the molecular graph encoder. Our proposed method outperforms the strong baseline method Graph2Edits by 1.6% and 3.2% in terms of the top-1 accuracy and top-1 round-trip accuracy on the USPTO-50K dataset, respectively, and also achieves a new state-of-the-art performance among semi-template-based methods on the USPTO-FULL dataset. The source code of this work and associated trained models are available at the KGCL GitHub : https://github.com/mrzhaodekui/KGCL.

1 Introduction

Retrosynthesis prediction is an indispensable strategy for designing drugs, from small molecules to complex natural products[Gothard *et al.*, 2012]. AI-based models for retrosynthesis prediction has achieved great progresses, which are divided into three classes: template-based, template-free, and semi-template-based methods. Template-based methods cannot predict reactions beyond the available templates, despite their strong interpretability [Segler and Waller, 2017; Dai *et al.*, 2019; Chen and Jung, 2021; Xie *et al.*, 2023]. Template-free methods do not rely on predefined templates [Tetko *et al.*, 2020; Seo *et al.*, 2021], but the reactants generated by these methods are probably chemically invalid and lack interpretability[Wan *et al.*, 2022; Tu and Coley, 2022; Yao *et al.*, 2024; Han *et al.*, 2024]. To balance the dependence on templates and interpretability of the prediction process, semi-template-based methods were proposed, which perform retrosynthesis in two sequential steps: (1) breaking the molecule by identifying a reaction center, and (2) transforming the resulting fragments into potential reactants[Sacha *et al.*, 2021].

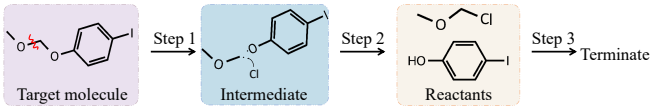


Figure 1: An example of retrosynthesis prediction based on molecular graph editing: Given the target molecule 1-iodo-4-(methoxymethoxy) benzene, the first edit predicted by the model is the removal of the C:3–O:4 bond. The second edit predicted by the model is the attachment of 'Cl' to C:3. Finally, a termination symbol indicates the completion of the retrosynthesis process.

The retrosynthesis prediction method based on molecular graph editing, which is inspired by the arrow-pushing method in reaction mechanism description, is one of the most representative semi-template-based strategies. As shown in Figure 1, the molecular graph editing methods formulate retrosynthesis as a sequential modification process on the product graph, which is guided by simplified reaction mechanisms. However, existing methods typically treat individual atoms in molecules as independent functional units when learning molecular graph representations. They ignore functional groups, which are substituents or moieties in molecules that cause the molecule's characteristic chemical reactions, thus limiting the predictive performance of the models[Zhong *et al.*, 2023].

To address this issue, this paper incorporates the functional group knowledge from the chemical knowledge graph into the embeddings of molecular graphs. In details, we first extract functional group knowledge related to the target molecule from the knowledge graph, and then employ an attention mechanism to integrate the functional group knowledge into the atomic representation of the molecular graph. Such that the learned molecular embeddings become more discriminative and interpretable.

Furthermore, to enhance the accuracy of molecular graph representations, this paper introduces a contrastive learning

*Corresponding Author

strategy based on molecular graphs. The primary challenge is how to construct the rational positive and negative samples for the target molecular graph. Traditional random graph augmentation strategies are not suitable due to unique chemical constraints [Fang *et al.*, 2023]. To address this issue, we use graph edit operations predicted by the model as perturbations to generate positive and negative samples for the target molecular graph.

The main contributions are summarized as follows:

- We propose a strategy to incorporate the chemical functional group knowledge from knowledge graph into retrosynthesis prediction based on molecular graph editing. To the best of our knowledge, this is the first study to apply the chemical knowledge graph to retrosynthesis prediction within the context of molecular graph editing.
- We present a new molecular graph contrastive learning method. This method generates positive and negative samples for the target molecular graph using molecular graph edit operations predicted by the model, thereby preserving the semantic information of molecular graphs to the greatest extent.
- Experiments on two benchmark datasets demonstrate that our proposed model KGCL consistently outperforms the compared baseline methods and achieve new state-of-the-art performance among the semi-template-based methods.

2 Related Work

2.1 Knowledge-aware Semi-template-based Retrosynthesis Prediction

A series of semi-template-based retrosynthesis models were proposed in [Shi *et al.*, 2020; Somnath *et al.*, 2021; Wang *et al.*, 2021; Sacha *et al.*, 2021; Zhong *et al.*, 2023]. Although achieved some progresses, these models did not consider the chemical knowledge when learning molecular representation. Thus the prediction performance of these models needs to be improved. Afterward, the models in [Chen *et al.*, 2023; Liu *et al.*, 2024] embedded the chemical knowledge, such as reaction center types and chemical synthesis rules, into the retrosynthesis systems to improve the accuracy and interpretability. But in these models, the knowledge was predefined and lacked flexibility. In the present paper, the knowledge is from the chemical knowledge graph, not from predefined rules. The knowledge graph has achieved great success to predict molecular properties[Fang *et al.*, 2023], but not yet for retrosynthesis predictions. In the two scenarios, the fusion strategies of the chemical domain knowledge are different.

2.2 Molecular Graph Contrastive Learning

Inspired by the success of contrastive learning in image and language domains, it was introduced to molecular learning and was shown to be effective [Chen *et al.*, 2024]. Some researchers generated positive and negative samples by random perturbations for data augmentation [Luo *et al.*, 2023; Zheng *et al.*, 2023; Sun *et al.*, 2022]. However, these methods inevitably introduce variance in critical semantic information, which can mislead contrastive learning [You *et al.*, 2020; Sun *et al.*, 2021a]. To address this issue, some works predefined some domain knowledge and then utilized them to generate two augmented views [Sun *et al.*, 2021a; Kim *et al.*, 2023]. In the present paper, we need not predefine any domain knowledge, but directly utilize the graph edits predicted by the model to construct positive and negative samples.

3 Proposed Method

The overall framework of the KGCL model is shown in Figure 2, and the details are in the following subsections.

3.1 Problem Definition

A molecule $\mathcal{M}$ is represented as a graph $\mathcal{G} = (\mathcal{A}, \mathcal{B})$ in retrosynthesis prediction based on molecular graph editing. $\mathcal{A}$ is the set of atoms and $\mathcal{B}$ is the set of bonds in the molecule, where each atom corresponds to a node and each bond corresponds to an edge. The graph edit set includes an atom-level edit set E_a, a bond-level edit set E_b and a molecule-level edit set E_g. Specifically, E_a includes (1) atom modification and (2) attaching a leaving group to the atom, E_b includes (1) bond modification and (2) bond deletion, and edits in E_g determine whether the graph editing process has been completed. The task of the retrosynthesis prediction model is, given a target product molecule $\mathcal{M}_p$ and its graph $\mathcal{G}_p$, to predict a sequence of graph edits $(e_u^1, \cdots, e_u^l, \cdots, e_u^L)$ with length L and $u \in \{a, b, g\}$ (i.e., $e_u^l \in E_a \cup E_b \cup E_g$), which modifies $\mathcal{G}_p$ sequentially, until the graphs $\mathcal{G}_{\text{reactant}}$ of the reactant molecules, which could potentially synthesize the product, are obtained. This process can be formally described as follows:

$$\mathcal{G}_{\text{reactant}} = e_u^L \circ \cdots \circ e_u^l \circ \cdots \circ e_u^1(\mathcal{G}_p), \quad \mathcal{G}_{\text{reactant}} \in \mathcal{R} \quad (1)$$

where $\mathcal{R}$ denotes the set of all possible reactant graphs. The priority order of these operations is the same as in [Zhong *et al.*, 2023].

3.2 Knowledge-based Molecular Graph Enhancement

In KGCL, the i-th atom and the undirected bond between the i-th and j-th atoms are initially encoded as $h_i = (h_{i,1}, \cdots h_{i,n_a})$ with dimension n_a and $h_{ij} = (h_{ij,1}, \cdots, h_{ij,n_b})$ with dimension n_b, respectively, for $1 \leq i, j \leq |\mathcal{A}|$. The initial features of atoms and bonds are predefined in terms of chemical properties. More details can be found in [Zhong *et al.*, 2023].

A functional group is a group of atoms in a molecule with distinctive chemical properties, regardless of the other atoms in the molecule. To enhance the accuracy of molecular learning, we integrate the functional group knowledge into the representations of molecular graphs.

This study only focuses on the functional groups that are actually present in the target molecule to reduce the noises introduced by irrelevant functional groups. Specifically, given a product molecule $\mathcal{M}_p$ and the predefined set of functional groups $F = \{f_1, \cdots, f_i, \cdots, f_n\}$, where n is the number of predefine functional groups, we first determine whether each f_i is a substructure of the molecule

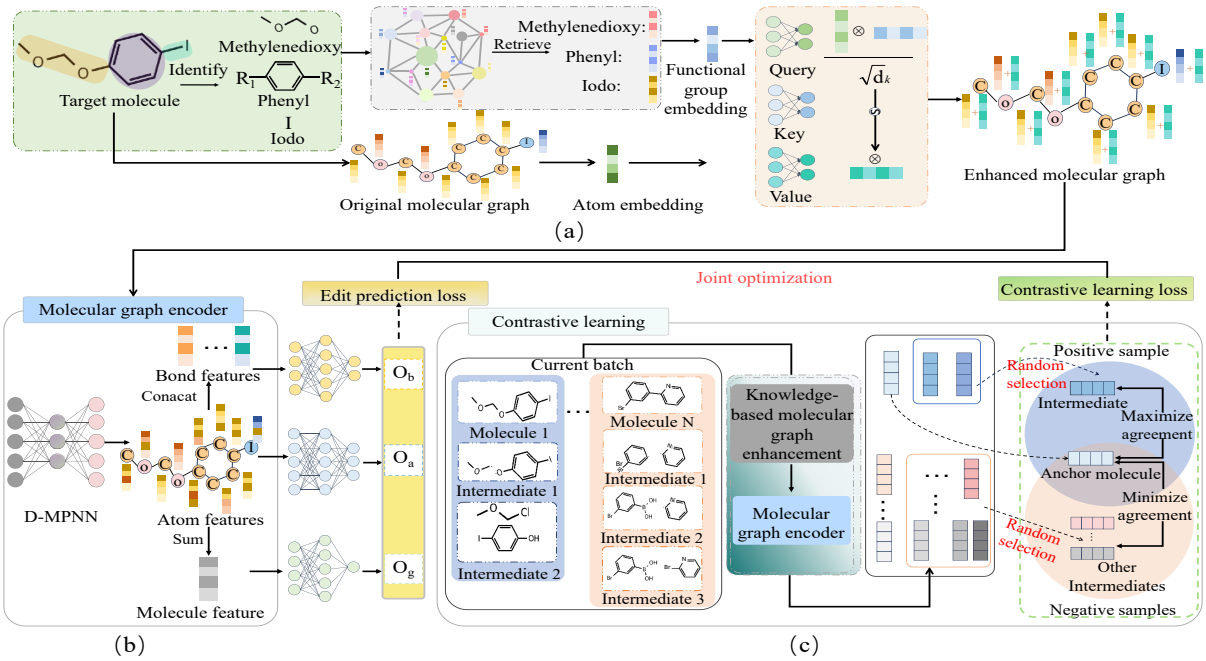


Figure 2: Overview of the proposed KGCL model. (a) Knowledge-based molecular graph enhancement, (b) The molecular graph encoder and the predictors for molecular graph edits, and (c) Molecular graph contrastive learning module.

$\mathcal{M}_p$ using the substructure matching algorithm. Therefore, we obtain the set of functional groups within $\mathcal{M}_p$, denoted as $F_p = \{f_{p_1}, \cdots, f_{p_j}, \cdots, f_{p_m}\}$, where m is the number of the functional groups in $\mathcal{M}_p$. Subsequently, the embeddings of these functional groups $H_{\text{func}} = [h^T_{\text{func},1}, \cdots, h^T_{\text{func},j}, \cdots, h^T_{\text{func},m}]^T$, where $h_{\text{func},j}$ is the embedding vector of f_{p_j}, are retrieved from the chemical knowledge graph [Fang *et al.*, 2023]. Finally, these functional group embeddings are incorporated into the representation of the molecular graph using an attention mechanism.

Specifically, the embeddings of the atoms in the original molecular graph are used as queries to attend to all functional groups. The knowledge-enhanced embedding $\tilde{h}_i$ for the i-th atom is computed as follows:

$$\text{Attention}(h_i, H_{\text{func}}) = \text{softmax}(\frac{h_i \cdot H^T_{\text{func}}}{\sqrt{d_f}}), \tag{2}$$

$$\tilde{h}_i = h_i + \text{Attention}(h_i, H_{\text{func}}) \cdot H_{\text{func}}, \tag{3}$$

where d_f is the dimension of the functional group embedding. Next, the embeddings of bonds are enhanced and injected direction information as follows:

$$\tilde{h}_{ij} = \tilde{h}_i \oplus h_{ij}, \tag{4}$$

$$\tilde{h}_{ji} = \tilde{h}_j \oplus h_{ij}, \tag{5}$$

where $\oplus$ denotes the concatenation operation. Finally, the knowledge-enhanced molecular graph is denoted as

$$\tilde{H}_g = (\{\tilde{h}_i | 1 \le i \le |\mathcal{A}|\}, \{\tilde{h}_{ij} | 1 \le i, j \le |\mathcal{A}|\}). \tag{6}$$

3.3 The Molecular Graph Encoder

We use the directed message passing neural network (D-MPNN) [Gilmer *et al.*, 2017] to encode the molecular graph. At the $(t+1)$-th iteration, hidden representation $\tilde{h}^{(t+1)}_{ij}$ of each edge is updated based on messages $m^{(t+1)}_{ij}$ according to

$$m^{(t+1)}_{ij} = \sum_{k \in N(i) \setminus j} M_t(\tilde{h}_i, \tilde{h}_k, \tilde{h}^{(t)}_{ik}), \tag{7}$$

$$\tilde{h}^{(t+1)}_{ij} = U_t(\tilde{h}^{(t)}_{ij}, m^{(t+1)}_{ij}), \tag{8}$$

$$\tilde{h}^{(0)}_{ik} = W_{\text{input}}(\tilde{h}_{ik}), \tag{9}$$

where $\tilde{h}_{ik}$ is the initial feature of the edge from the node i to the k in the knowledge-enhanced molecular graph, W_{input} is the learnable weight matrix, $N(i)$ is the neighbors of the i-th vertex, and m_{ij} is the message feature vector from the node i to the node j. M_t and U_t are the message and edge update function:

$$M_t(\tilde{h}_i, \tilde{h}_j, \tilde{h}^{(t)}_{ij}) = \tilde{h}^{(t)}_{ij}, \tag{10}$$

$$U_t(\tilde{h}^{(t)}_{ij}, m^{(t+1)}_{ij}) = GRU(\tilde{h}^{(t)}_{ij} + m^{(t+1)}_{ij}). \tag{11}$$

After T iterations, the embedding of the i-th vertex is updated as follows:

$$\tilde{h}_i = \sigma_r(W_{\text{output}}(\tilde{h}_i \oplus \sum_{j \in N(i)} \tilde{h}^{(T)}_{ji}) + c), \tag{12}$$

where σ_r is the SELU activation function, W_{output} is the learnable weight matrix and c is the bias of the fully connected network.

3.4 The Generation of Edit Sequences

The edit sequence $(e_u^1, \cdots, e_u^l, \cdots, e_u^L)$ for the target molecular graph is generated in an autoregressive manner [Zhong *et al.*, 2023], in which three MLPs are used to predict the atom-level, the bond-level and the graph-level edits, respectively. At the l-th step, to strengthen the relationship between the current editing operation and previous ones, the atom embeddings obtained from the D-MPNN encoder are updated by:

$$\tilde{h}_i^{(l)} = \sigma_r(W_p \tilde{h}_i^{(l-1)} + W_c \tilde{h}_i^{(l)}), \tag{13}$$

where $\tilde{h}_i^{(0)} = \tilde{h}_i$ defined in (12), W_p and W_c are the learnable weight matrices, and σ_r is the SELU activation function. After the atom features are updated, the bond features are recomputed using the features of the atoms at its two ends, and the molecular features are obtained by summing the hidden representation of all its atoms as follows:

$$\tilde{h}_{ij}^{(l)} = \tilde{h}_i^{(l)} \oplus \tilde{h}_j^{(l)}, \tag{14}$$

$$\tilde{h}_{\tilde{\mathcal{G}}}^{(l)} = \sum_{i \in \tilde{\mathcal{G}}^{(l)}} \tilde{h}_i^{(l)}, \tag{15}$$

where $\oplus$ denotes the concatenation operation.

At last, the logits $o_{(a,i)}^{(l)}$, $o_{(b,ij)}^{(l)}$ and $o_g^{(l)}$ corresponding to the atom-level edit $a \in E_a$, bond-level edit $b \in E_b$, and graph-level edit at the l-th step are predicted as follows:

$$o_{(i,a)}^{(l)} = W_2(\sigma_r(W_1 \tilde{h}_i^{(l)} + c_1) + c_2), \tag{16}$$

$$o_{(ij,b)}^{(l)} = W_4(\sigma_r(W_3 \tilde{h}_{ij}^{(l)} + c_3) + c_4), \tag{17}$$

$$o_g^{(l)} = W_6(\sigma_r(W_5 \tilde{h}_{\tilde{\mathcal{G}}}^{(l)} + c_5) + c_6), \tag{18}$$

where W_i is the weight matrix, and c_i is the corresponding bias for $1 \leq i \leq 6$, and σ_r is the SELU activation function.

3.5 Molecular Graph Contrastive Learning Based on Edit Sequences

In order to enhance the representation learning ability of the encoder, we propose a novel molecular graph contrastive learning strategy. One main challenge in contrastive learning is to generate rational positive and negative samples:

- The anchor sample should be semantically similar to the positive sample while far from the negative samples;
- In our context, the molecules corresponding to the positive and negative samples should satisfy the chemical property constraints.

To address this issue, we leverage predicted edit sequences as augmentation operations to construct positive and negative samples. Specifically, let $S_{\text{batch}} = \{\mathcal{M}_1, \cdots, \mathcal{M}_i, \cdots, \mathcal{M}_N\}$ denote the set of product molecules in the current training batch. For each molecule $\mathcal{M}_i$, the predicted edit sequence is $(e^{i_1}, e^{i_2}, \cdots, e^{i_{L_i}})$, where L_i is the length of current edit sequence for M_i. This sequence generates a set of intermediates $S_i^+ = \{\mathcal{M}_i^l = e^{i_l} \circ \cdots \circ e^{i_1}(\mathcal{M}_i) | 1 \leq l \leq L_i\}$. We randomly select one $\mathcal{M}_i^+$ from S_i^+ as the positive sample of $\mathcal{M}_i$. The set of negative samples for $\mathcal{M}_i$ is defined as $S_i^- = \cup_{j \neq i}\{\mathcal{M}_j^1, \mathcal{M}_j^2, \cdots, \mathcal{M}_j^{L_j}\}$, which includes intermediates generated from all other molecules in the batch.

In this study, we adopt a variant of InfoNCE loss as the contrastive loss, which assigns dynamic Gaussian weights to negative samples[Wu *et al.*, 2024]. The contrastive loss function is defined as:

$$\mathcal{L}_{\cos} = -\frac{1}{N}\sum_{i=1}^{N}(\log(\exp\frac{\tilde{h}_{\tilde{\mathcal{G}}_i} \cdot \tilde{h}_{\tilde{\mathcal{G}}_i^+}}{\tau}) - (\log(\exp\frac{\tilde{h}_{\tilde{\mathcal{G}}_i} \cdot \tilde{h}_{\tilde{\mathcal{G}}_i^+}}{\tau}) + (\sum_{M_j^- \in S_i^-} \exp\frac{w_j \cdot (\tilde{h}_{\tilde{\mathcal{G}}_i} \cdot \tilde{h}_{\tilde{\mathcal{G}}_j^-})}{\tau}))), \tag{19}$$

where $\tilde{h}_{\tilde{\mathcal{G}}_i}$, $\tilde{h}_{\tilde{\mathcal{G}}_i^+}$ and $\tilde{h}_{\tilde{\mathcal{G}}_j^-}$ are molecular graph embedding for the molecule M_i, its positive sample M_i^+ and its negative sample M_j^- as defined in (15), τ is temperature coefficient and w_j is the weight of the negative sample M_j^-, which is dynamically computed based on the similarity between the anchor sample M_i and the negative sample M_j^- as follows:

$$w_j = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{(\tilde{h}_{\tilde{\mathcal{G}}_i} \cdot \tilde{h}_{\tilde{\mathcal{G}}_j^-}) - \mu}{\sigma})^2), \tag{20}$$

where μ and σ, two hyperparameters, are the mean and variance, respectively. Samples closer to μ have larger weights. The smaller the value of σ is, the more pronounced the weight differences between samples are.

3.6 Overall Loss Function for Model Training

We utilize the following cross-entropy loss, referred to as the edit prediction loss, to measure the difference between the ground truth labels and the predicted labels:

$$\mathcal{L}_{\text{edits}} = -\frac{1}{L}\sum_{l=1}^{L}\Bigg(\sum_{\mathcal{G}_m \in batch}\Bigg(\sum_{a \in E_a} y_i^{(l)} \log(o_{(a,i)}^{(l)}) + \sum_{b \in E_b} y_{ij}^{(l)} \log(o_{(b,ij)}^{(l)}) + y_{\tilde{\mathcal{G}}}^{(l)} \log(o_{\tilde{\mathcal{G}}}^{(l)})\Bigg)\Bigg), \tag{21}$$

where L is the length of the graph edit sequence, $\mathcal{G}_m$ is an intermediate in the current batch during the training process. $y_i^{(l)}$, $y_{ij}^{(l)}$ and $y_{\tilde{\mathcal{G}}}^{(l)}$ are the ground-truth labels for the atom edit $a \in E_a$, the bond edit $b \in E_b$ and the molecular graph edit at the l-th step. $o_{(a,i)}^{(l)}$, $o_{(b,ij)}^{(l)}$ and $o_{\tilde{\mathcal{G}}}^{(l)}$ are the output scores of the model for the corresponding edits.

The overall loss is defined as the weighted combination of the edit prediction loss and the contrastive learning loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{edits}} + \lambda * \mathcal{L}_{\cos}, \tag{22}$$

where λ is a hyperparameter that controls the weight of the contrastive loss.

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments are conducted on public benchmark datasets USPTO-50K [Schneider *et al.*, 2016] and USPTO-FULL [Dai *et al.*, 2019]. The USPTO-50K is a high-quality dataset that contains 50,016 reactions with the correct atom-mapping. It is divided into the training, validation, and test sets according to the partitioning scheme as [Coley *et al.*, 2017], with each set containing 40k, 5k, and 5k reactions, respectively. Compared with the USPTO-50K dataset, the USPTO-FULL dataset has higher coverage and diversity, containing 1000k unfiltered chemical reactions. We adopt the same splitting method as [Dai *et al.*, 2019] and divide it into 800k, 100k, and 100k reactions as the training set, validation set, and test set respectively. For fair comparisons, the extraction of graph editing operations and other data preprocessing are the same as those in [Zhong *et al.*, 2023].

Implementation details. The KGCL model uses the Adam optimizer for gradient descent training. For the USPTO-50K dataset, the initial learning rate is set to 0.001 (0.0001 for the USPTO-FULL dataset), and a polynomial decay learning rate scheduler is applied. When the improvement in accuracy on the validation set is less than a predefined threshold 0.01 within 5 consecutive epochs, it is considered to have reached a plateau (or peak). At this point, the learning rate is reduced by a decay factor 0.8. The hidden dimension of the encoder is set to 256, the number of iterations of message passing is 10, and node embeddings are dropped with the probability 0.15. We use three MLPs with a hidden dimension 512 and a dropout rate of 0.2 to predict the initial edit scores. The model was trained for 200 epochs using a batch size of 256. All experiments were conducted on a single NVIDIA RTX A6000 GPU.

Evaluation and baselines. We employ the top-k exact match, round-trip and MaxFrag accuracy as the metrics to measure the retrosynthesis performance. The top-k exact match accuracy is calculated by comparing the standard Simplified Molecular Input Line Entry System (SMILES) of predicted reactants with the ground-truths in the dataset [Weininger, 1988]. We report the top-k accuracy where k=1, 3, 5, 10 and 50. The round-trip accuracy is computed by evaluating the similarity between the ground-truth product and the product generated by a forward-synthesis model using the predicted reactants [Schwaller *et al.*, 2020]. We utilize Molecular Transformer (MT) [Schwaller *et al.*, 2019] as a forward-synthesis prediction model to compute the round-trip accuracy. The MaxFrag accuracy is defined as the exact match between the largest fragment of the predicted reactant and the ground-truth in the dataset [Tetko *et al.*, 2020].

We compared our model with representative baseline models in recent years, include:

- template-based methods: Retrosim[Coley *et al.*, 2017], Neuralsym [Segler and Waller, 2017], GLN [Dai *et al.*, 2019], LocalRetro [Chen and Jung, 2021], RetroKNN[Xie *et al.*, 2023];
- Tempate-free methods: Aug.Transformer [Tetko *et al.*, 2020], GTA[Seo *et al.*, 2021], Dual-TF [Sun *et al.*, 2021b], Retroformer [Wan *et al.*, 2022], Graph2SMILES [Tu and Coley, 2022], NAG2G [Yao *et al.*, 2024];
- Semi-template-based methods: MEGAN [Sacha *et al.*, 2021], RetroPrime [Wang *et al.*, 2021], GraphRetro [Somnath *et al.*, 2021], G2Retro [Chen *et al.*, 2023], Graph2Edits [Zhong *et al.*, 2023], MARS [Liu *et al.*, 2024].

4.2 Main Results

The comparison of model performance, in terms of the top-k exact match accuracy, on the USPTO-50K dataset is presented in Tables 1 and 2. When the reaction class is unknown, the KGCL model achieved a top-1 accuracy of 56.3%. In more detail, KGCL achieves state-of-the-art performance among all semi-template-based methods for all values of k and surpasses the strong baseline Graph2Edits by margins of 1.2% and 1.9% in top-1 and top-10 accuracy, respectively. In addition, the performance of KGCL outperforms those of all template-free methods and is better than those of most template-based methods. Even compared with RetroKNN, a strong baseline for template-based methods, KGCL also shows strong competitiveness. When the reaction type is known, KGCL also outperforms all other semi-template-based methods for all k values and improves the top-1 accuracy by 1.6% compared to the strong baseline Graph2Edits.

Model	Top-k accuracy (%)				
	k=1	3	5	10	50
Template-Based Methods					
GLN	52.5	69.0	75.6	83.7	92.4
LocalRetro	53.4	77.5	85.9	92.4	97.7
RetroKNN	**57.2**	**78.9**	**86.4**	**92.7**	**98.1**
Template-Free Methods					
Retroformer	53.2	71.1	76.6	82.1	-
Graph2SMILES	52.9	66.5	70.0	72.9	-
Dual-TF	53.6	70.7	74.6	77.0	-
NAG2G	**55.1**	**76.9**	**83.4**	**89.9**	-
Semi-Template-Based Methods					
MEGAN	48.1	70.7	78.4	86.1	93.2
GraphRetro	53.7	68.3	72.2	75.5	-
MARS	54.6	76.4	83.3	88.5	-
G2Retro	54.1	74.1	81.2	86.7	-
Graph2Edits	55.1	77.3	83.4	89.4	92.7
KGCL(ours)	**56.3**	**78.1**	**85.4**	**91.3**	**96.1**

Table 1: Top-k exact match accuracy on USPTO-50K Dataset without reaction classes.

To evaluate the performance of our model in generating valid retrosynthetic suggestions and predicting the correct reaction class, we compared KGCL with baseline methods in terms of round-trip and MaxFrag accuracy. The comparative results on the USPTO-50K dataset are presented in Table 3. The top-1 round-trip accuracy of KGCL reached 89.1%, which is nearly comparable to the performance of LocalRetro and surpasses the strong baseline Graph2Edits by

3.2%. For other values k, KGCL also significantly outperforms Graph2Edits, demonstrating the validation of its predictive results. Furthermore, in terms of MaxFrag accuracy, KGCL achieves a top-1 accuracy of 60.8%, significantly outperforming all baseline methods.

Model	Top-k accuracy (%)				
	k=1	3	5	10	50
Template-Based Methods					
LocalRetro	63.9	86.8	92.4	96.3	97.9
GLN	64.2	79.1	85.2	90.0	93.2
RetroKNN	**66.7**	**88.2**	**93.6**	**96.6**	**98.4**
Template-Free Methods					
Retroformer	64.0	82.5	86.7	90.2	-
Dual-TF	65.7	81.9	84.7	85.9	-
NAG2G	**67.2**	**86.4**	**90.5**	**93.8**	-
Semi-Template-Based Methods					
MEGAN	60.7	82.0	87.5	91.6	95.3
GraphRetro	63.9	81.5	85.2	88.1	-
G^2Retro	63.6	83.6	88.4	91.5	-
MARS	66.2	85.8	90.2	92.9	-
Graph2Edits	67.1	87.5	91.5	93.8	94.6
KGCL(ours)	**68.7**	**87.9**	**92.0**	**94.5**	**95.8**

Table 2: Top-k exact match accuracy on USPTO-50K Dataset with reaction classes.

Category	Model	Top-k accuracy (%)				
		k=1	3	5	10	50
Round-Trip accuracy	Template-Based Methods					
	LocalRetro	**89.5**	**97.9**	**99.2**	-	-
	Semi-Template-Based Methods					
	MEGAN	82.0	89.9	91.7	94.0	96.4
	GraphRetro	86.0	89.9	90.7	91.4	91.6
	Graph2Edits	85.9	93.5	95.1	96.4	97.3
	KGCL(ours)	**89.1**	**97.5**	**98.7**	**99.5**	**99.9**
MaxFrag accuracy	Template-Based Methods					
	LocalRetro	**57.8**	**82.1**	**89.7**	**95.0**	**98.4**
	Template-Free Methods					
	Aug.Transformer	**58.5**	**73.0**	**85.4**	**90.0**	-
	Semi-Template-Based Methods					
	MEGAN	54.2	75.7	83.1	89.2	95.1
	Graph2Edits	59.2	80.1	86.1	91.3	93.1
	KGCL(ours)	**60.8**	**81.3**	**87.9**	**93.0**	**96.8**

Table 3: Top-k round-trip and MaxFrag accuracy on USPTO-50K dataset without reaction classes.

In order to verify the generalization ability and robustness of our method, we evaluated the model performance on the USPTO-FULL dataset, which has a larger number of reactions and more diverse reaction classes. The results are shown in Table 4. For fair comparison, we did not clean the USPTO-FULL dataset, including the test set. Despite the large amount of noises in the original dataset, our method still achieve considerable performance. Moreover, for k=1, 3, 5, 10, our model achieves state-of-the-art performance among all semi-template-based retrosynthesis methods. Additionally, KGCL also shows strong competitiveness compared with current leading template-free methods, which further highlights its potential for efficient search in more complex and diverse reaction spaces.

Model	Top-k Accuracy (%)			
	k=1	3	5	10
Template-Based Methods				
Retrosim	32.8	-	-	56.1
Neuralsym	35.8	-	-	60.8
LocalRetro	39.1	53.3	58.4	63.7
GLN	**39.3**	-	-	**63.7**
Template-Free Methods				
Graph2SMILES	45.7	-	-	63.4
Aug.Transformer*	46.2	–	-	73.3
GTA*	46.6	-	-	70.4
NAG2G*	**47.7**	**62.0**	**66.6**	**71.0**
Semi-Template-Based Methods				
MEGAN	33.6	-	-	63.9
RetroPrime*	44.1	59.1	62.8	68.5
Graph2Edits	44.0	60.9	66.8	72.5
KGCL(ours)	**44.1**	**61.4**	**67.3**	**73.0**

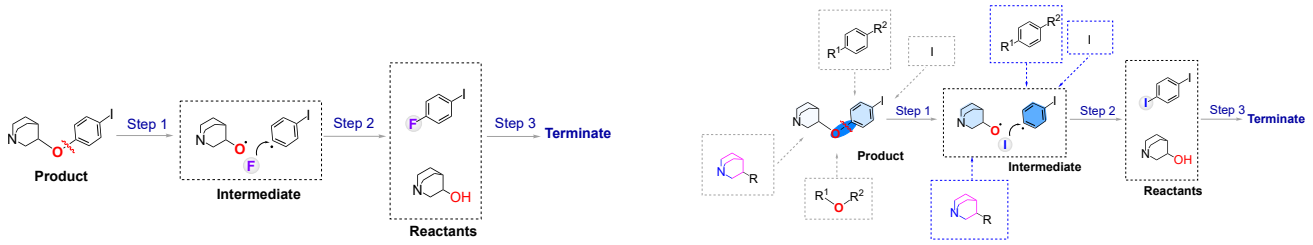
Table 4: Top-k exact match accuracy on USPTO-FULL dataset without reaction classes. * indicates that invalid reactions in the test set were removed.

4.3 Ablation Study

To investigate the contributions of different components, we conducted ablation experiments on the USPTO-50K dataset. The results are presented by comparing the top-1 exact match accuracy of different models on the USPTO-50K dataset without reaction classes, as shown in Table 5. Model I removes the functional group knowledge and contrastive learning strategy, and its accuracy is 55.1%. Model II adds the graph contrastive learning strategy into Model I, and the top-1 accuracy increases by 0.5% to 55.6%, which proves the effectiveness of the graph contrastive learning strategy. Similarly, Model III incorporates the functional group knowledge into Model I, and the top-1 accuracy increases by 0.4%, reaching 55.5%. The results show the importance of the functional group knowledge for the retrosynthesis prediction task. Model IV is a complete KGCL model. Its top-1 accuracy is improved by 1.2% compared to Model I, which exceeds the cumulative gains achieved by using the functional group knowledge and the graph contrastive learning strategy independently. This demonstrates the superior performance achieved when these two strategies work together. Therefore, every component is indispensable.

4.4 Case Study

To better understand the impact of the functional group knowledge on retrosynthetic predictions, we randomly selected a product molecule from the test set of the USPTO-50K dataset. As shown in Figure 3, in the first step, Graph2Edits and our model predicted correctly. In the second step,



(a) The retrosynthesis prediction process of Graph2Edits

(b) The retrosynthesis prediction process of our method

Figure 3: Investigation of importance of the functional group knowledge. Darker colors indicate higher weights of the functional groups.

our model still predicted correctly, but Graph2Edits made a wrong prediction. The following is a detailed analysis: All halogen elements possess oxidability due to the same outer electrons number, but with the increase of the electron layers, from F to I, electrophilicity is decreasing and nucleophilicity continue to increase. It is well known that iodine is not only a good nucleophile but also a good leaving group. Graph2Edits incorrectly predicted to attach the fluorine atom to the benzene ring in the second step. And our model identified the presence of phenyl and iodine functional groups in the intermediate. Thus it predicted to attach the iodine atom to the benzene ring. Our prediction is more reasonable: On one side, the iodobenzene radical intermediates has high reactivity and presents electron-deficient property for the iodine substituents, thus it is more suitable to react with the nucleophilic iodine atoms. On the other side, iodine atom has the maximum atomic radius among halogens, which makes the C-I bond easily broken. Therefore, the reactants predicted by our model can produce the target product more easily.

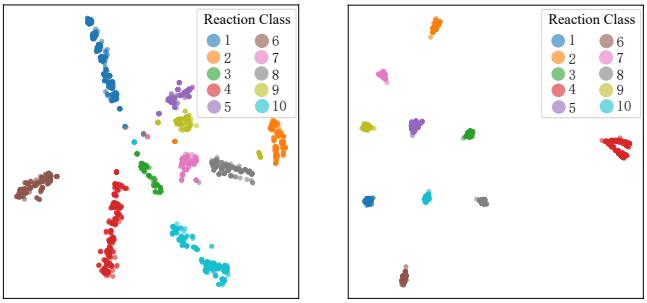
4.5 Visualization of Molecular Graph Embeddings

To evaluate the quality of the learned molecular embeddings, we visualize the embeddings of complex product molecules and their intermediates from different reaction classes, as shown in Figure 4. Specifically, we randomly selected a complex product molecule from each reaction class in the test set of the USPTO-50K dataset. The selected molecules have at least one of the following characteristics: multiple reaction centers, stereochemical features, or a long synthetic pathway [Herges, 1994; Zhong *et al.*, 2023]. We then used Graph2Edits and KGCL with beam width = 10 and beam size = 50 to generate graph edit sequences for these molecules and obtain their intermediates. Subsequently, the high-dimensional feature vectors of these molecules and their intermediates were reduced to a two-dimensional embedding space using the t-distributed stochastic neighbor embedding (t-SNE) [Van der Maaten and Hinton, 2008] method.

It can be seen from Figure 4a that the embeddings of product molecules and their intermediates generated by Graph2Edits present a dispersed distribution, and the boundaries of the intermediate embeddings of different target molecules are not sufficiently distinct. In contrast, KGCL (Fig. 4b) is able to effectively distinguish intermediate embeddings of different target molecules, and all intermediate embeddings of the same target molecule are tightly clustered together. The visualization results show that our model is able to capture differences between different molecular structures and maintain the similarity between different intermediates of the same molecule, which provides strong support to predict the graph edit sequences.

	Functional Group	Contrastive Learning	Top-1 (%)
I	×	×	55.1
II	×	✓	55.6
III	✓	×	55.5
IV	✓	✓	56.3

Table 5: Ablation study on the USPTO-50K Dataset without reaction classes.



(a) Graph2Edits

(b) Our method

Figure 4: t-SNE visualization of the embeddings for complex molecules from different reactions.

5 Conclusion

In the present paper, we improved the accuracy of retrosynthesis prediction by introducing the functional knowledge and the contrastive learning strategy. The experimental results on two benchmark datasets, USPTO-50K and USPTO-FULL, verified the effectiveness of our model KGCL. The success of KGCL depends on merging the domain knowledge from the chemical knowledge graph into the deep learning model in a rational way. In the future, we will further improve the prediction accuracy by introducing more domain knowledge, such as the electronic effect and steric effect, into the retrosynthesis prediction model.

Acknowledgements

This study was supported in part by the National Natural Science Foundation of China (Grant No. 62472082, 62277009).

References

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

[Chen *et al.*, 2023] Ziqi Chen, Oluwatosin R Ayinde, James R Fuchs, Huan Sun, and Xia Ning. G2retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6(1):102, 2023.

[Chen *et al.*, 2024] Jiayuan Chen, Kehan Guo, Zhen Liu, Olexandr Isayev, and Xiangliang Zhang. Uncertainty-aware yield prediction with multimodal molecular features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8274–8282, 2024.

[Coley *et al.*, 2017] 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.

[Dai *et al.*, 2019] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.

[Fang *et al.*, 2023] Yin Fang, Qiang Zhang, Ningyu Zhang, Zhuo Chen, Xiang Zhuang, Xin Shao, Xiaohui Fan, and Huajun Chen. Knowledge graph-enhanced molecular contrastive learning with functional prompt. *Nature Machine Intelligence*, 5(5):542–553, 2023.

[Gilmer *et al.*, 2017] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. PMLR, 2017.

[Gothard *et al.*, 2012] Chris M Gothard, Siowling Soh, Nosheen A Gothard, Bartlomiej Kowalczyk, Yanhu Wei, Bilge Baytekin, and Bartosz A Grzybowski. Rewiring chemistry: algorithmic discovery and experimental validation of one-pot reactions in the network of organic chemistry. *Angewandte Chemie International Edition*, 51(32):7922–7927, 2012.

[Han *et al.*, 2024] Yuqiang Han, Xiaoyang Xu, Chang-Yu Hsieh, Keyan Ding, Hongxia Xu, Renjun Xu, Tingjun Hou, Qiang Zhang, and Huajun Chen. Retrosynthesis prediction with an iterative string editing model. *Nature Communications*, 15(1):6404, 2024.

[Herges, 1994] Rainer Herges. Organizing principle of complex reactions and theory of coarctate transition states. *Angewandte Chemie International Edition in English*, 33(3):255–276, 1994.

[Kim *et al.*, 2023] Seojin Kim, Jaehyun Nam, Junsu Kim, Hankook Lee, Sungsoo Ahn, and Jinwoo Shin. Fragment-based multi-view molecular contrastive learning. In *Workshop on"Machine Learning for Materials"ICLR 2023*, 2023.

[Liu *et al.*, 2024] Jiahan Liu, Chaochao Yan, Yang Yu, Chan Lu, Junzhou Huang, Le Ou-Yang, and Peilin Zhao. Mars: a motif-based autoregressive model for retrosynthesis prediction. *Bioinformatics*, 40(3):btae115, 2024.

[Luo *et al.*, 2023] Zhenfei Luo, Yixiang Dong, Qinghua Zheng, Huan Liu, and Minnan Luo. Dual-channel graph contrastive learning for self-supervised graph-level representation learning. *Pattern Recognition*, 139:109448, 2023.

[Sacha *et al.*, 2021] Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.

[Schneider *et al.*, 2016] 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.

[Schwaller *et al.*, 2019] 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.

[Schwaller *et al.*, 2020] Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.

[Segler and Waller, 2017] 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.

[Seo *et al.*, 2021] Seung-Woo Seo, You Young Song, June Yong Yang, Seohui Bae, Hankook Lee, Jinwoo Shin, Sung Ju Hwang, and Eunho Yang. GTA: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 531–539, 2021.

[Shi *et al.*, 2020] Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *International conference on machine learning*, pages 8818–8827. PMLR, 2020.

[Somnath *et al.*, 2021] Vignesh Ram Somnath, Charlotte Bunne, Connor Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34:9405–9415, 2021.

[Sun *et al.*, 2021a] Mengying Sun, Jing Xing, Huijun Wang, Bin Chen, and Jiayu Zhou. MoCL: data-driven molecular fingerprint via knowledge-aware contrastive learning from molecular graph. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 3585–3594, 2021.

[Sun *et al.*, 2021b] Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34:10186–10194, 2021.

[Sun *et al.*, 2022] Ruoxi Sun, Hanjun Dai, and Adams Wei Yu. Does gnn pretraining help molecular representation? *Advances in Neural Information Processing Systems*, 35:12096–12109, 2022.

[Tetko *et al.*, 2020] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575, 2020.

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

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wan *et al.*, 2022] Yue Wan, Chang-Yu Hsieh, Ben Liao, and Shengyu Zhang. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *International Conference on Machine Learning*, pages 22475–22490. PMLR, 2022.

[Wang *et al.*, 2021] Xiaorui Wang, Yuquan Li, Jiezhong Qiu, Guangyong Chen, Huanxiang Liu, Benben Liao, Chang-Yu Hsieh, and Xiaojun Yao. Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845, 2021.

[Weininger, 1988] 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.

[Wu *et al.*, 2024] Junkang Wu, Jiawei Chen, Jiancan Wu, Wentao Shi, Xiang Wang, and Xiangnan He. Understanding contrastive learning via distributionally robust optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

[Xie *et al.*, 2023] Shufang Xie, Rui Yan, Junliang Guo, Yingce Xia, Lijun Wu, and Tao Qin. Retrosynthesis prediction with local template retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 5330–5338, 2023.

[Yao *et al.*, 2024] Lin Yao, Wentao Guo, Zhen Wang, Shang Xiang, Wentan Liu, and Guolin Ke. Node-aligned graph-to-graph: Elevating template-free deep learning approaches in single-step retrosynthesis. *JACS Au*, 4(3):992–1003, 2024.

[You *et al.*, 2020] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33:5812–5823, 2020.

[Zheng *et al.*, 2023] Zixi Zheng, Yanyan Tan, Hong Wang, Shengpeng Yu, Tianyu Liu, and Cheng Liang. Casangcl: pre-training and fine-tuning model based on cascaded attention network and graph contrastive learning for molecular property prediction. *Briefings in Bioinformatics*, 24(1):bbac566, 2023.

[Zhong *et al.*, 2023] Weihe Zhong, Ziduo Yang, and Calvin Yu-Chian Chen. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1):3009, 2023.