FusionRetro: Molecule Representation Fusion via In-Context Learning for Retrosynthetic Planning

Songtao Liu[1] **Zhengkai Tu**[*2] **Minkai Xu**[*3] **Zuobai Zhang**[*45]
Lu Lin[1] **Rex Ying**[6] **Jian Tang**[478] **Peilin Zhao**[9] **Dinghao Wu**[1]

Abstract

Retrosynthetic planning aims to devise a complete multi-step synthetic route from starting materials to a target molecule. Current strategies use a decoupled approach of single-step retrosynthesis models and search algorithms, taking only the product as the input to predict the reactants for each planning step and ignoring valuable context information along the synthetic route. In this work, we propose a novel framework that utilizes context information for improved retrosynthetic planning. We view synthetic routes as reaction graphs and propose to incorporate context through three principled steps: *encode* molecules into embeddings, *aggregate* information over routes, and *readout* to predict reactants. Our approach is the first attempt to utilize in-context learning for retrosynthesis prediction in retrosynthetic planning. The entire framework can be efficiently optimized in an end-to-end fashion and produce more practical and accurate predictions. Comprehensive experiments demonstrate that by fusing in the context information over routes, our model significantly improves the performance of retrosynthetic planning over baselines that are not context-aware, especially for long synthetic routes. Code is available at https://github.com/SongtaoLiu0823/FusionRetro.

1. Introduction

Retrosynthetic planning is a fundamental problem in organic chemistry (Coley et al., 2018; Genheden et al., 2020).

[*]Equal contribution [1]Pennsylvania State University [2]Massachusetts Institute of Technology [3]Stanford University [4]Mila - Québec AI Institute [5]Université de Montréal [6]Yale University [7]HEC Montréal [8]CIFAR AI Chair [9]Tencent AI Lab. Correspondence to: Songtao Liu <skl5761@psu.edu>.
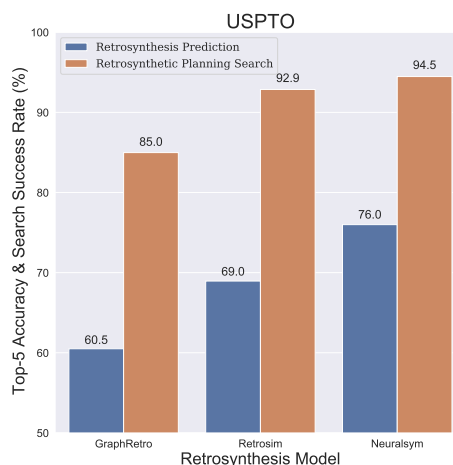
Figure 1. Performance of retrosynthesis prediction and multi-step planning on USPTO dataset. We report the search success rate of retrosynthesis models combined with Retro* at the limit of 500 calls and 5 expansions. The search success rate is much higher than the accuracy of the top-5 retrosynthesis prediction.

The goal of retrosynthetic planning is to find a viable set of starting materials and a sequence of reactions, that lead to a given target molecule. It is crucial for process chemistry, which aims to design efficient routes to synthesize desired target products at a low cost, as well as for materials and molecule discoveries that are contingent on the targets being synthesizable. In the past few years, with the advancement in deep learning, there has been increasing interest in applying machine learning to retrosynthetic planning, a sub-topic of Computer-Aided Synthesis Planning (CASP).

Existing strategies (Segler et al., 2018; Kishimoto et al., 2019; Chen et al., 2020; Lin et al., 2020; Schwaller et al., 2020; Kim et al., 2021; Xie et al., 2022; Yu et al., 2022) generally model retrosynthetic planning as a search problem. In a typical formulation, the synthetic route is treated as a tree or a graph, and the molecules as nodes. Starting from the target as the root node, these approaches employ some (possibly learned) search algorithms to select the most promising node to expand, and then expand it into reaction precursors with a one-step retrosynthesis model, until a viable route is found in which all the leaf nodes are commercially available.

It was not until recently that the evaluation criteria for multi-step search have somewhat converged to a few. One of the most heavily used metrics is the success rate of finding a viable route given an iteration limit (generally up to 500). However, the search success rate is overly lenient without checking whether the searched set of starting materials can go through a sequence of reactions to synthesize the target molecule at all. This is especially problematic for the targets requiring long routes to synthesize, in which case the errors can multiply. As a quantitative illustration in Figure 1, when we combine existing one-step models with top-5 accuracies between 60 and 80 percents with Retro* (Chen et al., 2020), an established search algorithm, the search success rates easily reach over 85 and 94 percents respectively. This is counterintuitive as we would expect the route to be less likely to succeed as more synthesis steps are added, which also throws concerns about the quality of proposed routes that the multi-step planner deems as "successful".

In this work, we therefore introduce the set-wise exact match of proposed starting materials to the ground truth as an alternative metric that better reflects reality. The underlying assumption is that if we can get the set of building blocks right, recovering all the reactions that ultimately lead to the target is a much easier process, possibly with the help of powerful reaction outcome predictors that can have a high accuracy of more than 90% (Irwin et al., 2022; Tetko et al., 2020). We construct a new benchmark with 58,099 synthetic routes retrieved from the public USPTO dataset for evaluation, in which we study the performance of multiple single-step retrosynthesis models in the context of multi-step planning, an important comparison that has not yet been done to the best of our knowledge. In addition, with our new framework of evaluation, it is now possible to consider and thereby improve the performance of all pieces in the CASP workflow in an integrated and holistic manner.

Under this view, it becomes immediately apparent that a missed opportunity by previous work is the explicit modeling of the contextual information of in-context reactions along the partial synthetic routes preceding any given node, which we subsequently explore. We propose a novel and principled context-aware model by fusing in the context embeddings, named **FusionRetro**, which is the first attempt to exploit in-context learning (Min et al., 2022) for retrosynthesis prediction in retrosynthetic planning. Specifically, we view the synthetic routes as reaction graphs and formulate our model as an end-to-end framework which: 1) **encodes** molecules on the synthetic routes into embeddings through molecule encoders; 2) **aggregates** the embeddings of molecules on the synthetic route (reaction graph) by message passing and fuses in the representations of informative contexts; and 3) **readouts** to predict the reactants on the current retrosynthetic step based on both the product and context representations learned in the previous stage.

Extensive experimental results on retrosynthetic planning tasks show that FusionRetro can achieve significantly better performance over template-free baselines, with up to a 6% improvement in top-1 test accuracy. The surprisingly superior performance demonstrates the effectiveness of exploiting the context information and opens up room for future research in this direction.

Our contribution can be summarized as follows:

- We introduce a new evaluation protocol for assessing the performance of single-step retrosynthesis models in the context of multi-step planning, and for this purpose, we curate a new benchmark dataset. Our empirical analysis confirms the pivotal role of single-step accuracy in multi-step planning.

- We propose a novel fusion framework that enables single-step models to leverage the contextual information in the reaction graph. Our method serves as the pioneering effort to exploit in-context learning (Dong et al., 2022) for solving scientific problems, which led to impressive success in recent large language models such as ChatGPT (Brown et al., 2020).

- Extensive experimental results demonstrate that our proposed module can enhance the performance of the baseline model noticeably, providing insightful guidance for future research in this direction.

2. Related Work

Single-step Retrosynthesis Model. Existing machine learning approaches for single-step retrosynthesis prediction can be classified into template-based and template-free models based on whether they rely on the use of reaction templates.

Template-based algorithms (Chen et al., 2020; Coley et al., 2017; Dai et al., 2019; Segler & Waller, 2017; Chen & Jung, 2021; Seidl et al., 2021) first extract these patterns from the training data, and then formulate the task as template classification or template retrieval. One of the intrinsic limitations of template methods is the need to find the right level of specificity for template definition so that they can capture sufficient chemical information without being overly specific to any reaction. As a remediation, researchers have come up with template-free methods, which have become more and more popular recently.

Template-free approaches generally use an end-to-end translation-based (Liu et al., 2017; Zheng et al., 2019; Chen et al., 2019; Karpov et al., 2019; Sun et al., 2021) or a graph-edit based formulation (Sacha et al., 2021). The former models the product-to-reactants transformation as a sequence-to-sequence task by representing molecules with

SMILES string, and the latter as a sequence of graph edits to atoms and bonds.

A special family of template-free methods, which are commonly referred to as semi-template-based methods (Shi et al., 2020; Yan et al., 2020; Somnath et al., 2021), adopts a two-stage formulation to first identify the reaction center(s). The target is subsequently broken into several disconnected subgraphs (i.e., the synthons), based on which the full molecule structures of reactants are recovered either by attaching the leaving group (Somnath et al., 2021) or by generative modeling (Shi et al., 2020; Yan et al., 2020). For a more comprehensive understanding of the retrosynthesis literature, readers are encouraged to refer to the survey paper (Meng et al., 2023).

Search Algorithm in Retrosynthetic Planning. Existing deep learning-based CASP models treat retrosynthetic planning as a search problem, which can be classified into Monte Carlo Tree Search (MCTS) (Segler et al., 2018; Hong et al., 2021), Proof-Number Search (PNS) (Kishimoto et al., 2019), A*-like Search (Chen et al., 2020; Han et al., 2022; Xie et al., 2022), and Reinforcement Learning (RL) based Search (Yu et al., 2022). Segler et al. (2018) integrates MCTS with policy networks to guide multi-step planning. Drawing inspiration from search techniques in two-player zero-sum games, DFPN-E (Kishimoto et al., 2019) combines Depth-First Proof-Number (DFPN) with Heuristic Edge Initialization for chemical synthesis planning. Retro* (Chen et al., 2020) introduces a neural-based A*-like algorithm to estimate solution costs and select the most promising one. GRASP (Yu et al., 2022) leverages reinforcement learning to guide the search process. Both GNN-Retro (Han et al., 2022) and RetroGraph (Xie et al., 2022) employ graph neural networks (Kipf & Welling, 2017) to aggregate information from the synthetic route, thereby enabling more accurate estimation of costs in Retro*. All works thus far, however, treat the selection policy and the expansion policy (i.e., the single-step model) as two disjoint pieces. Usage of context information of partially explored synthesis trees is non-existent in the single-step predictor, and to a minimal extent in the search phase in the form of some cost functions that are updated as planning proceeds.

In contrast, our work *explicitly* integrates reactions along the synthetic routes as in-context examples into our single-step model. We achieve this by fusing in the product embedding directly into the model inputs. Compared with context-aware A*-like search algorithms (Han et al., 2022; Xie et al., 2022), our proposed approach, focusing on retrosynthesis prediction, is a modular framework comprising encoding, fusion, and readout components. Our fusion leverages in-context learning to maximize the use of in-context reactions. Importantly, it is not limited to GNNs alone and can incorporate various aggregation methodologies, such as Transformer (Vaswani et al., 2017) and Graph Transformer (Ying et al., 2021). This framework lays a solid foundation for future explorations in the design of retrosynthesis models for retrosynthetic planning, specifically targeting three key aspects: encoding, fusion, and readout modules.

Evaluation of Retrosynthetic Planning. The de facto standard for evaluating single-step retrosynthesis models has been the top-k accuracy, or whether the ground truth reactants appear in the top-k suggestions. Alternatives such as accuracy for the largest predicted fragment (Tetko et al., 2020) and round-trip accuracy based on how likely the proposed reactants can lead to the product (Schwaller et al., 2020) have been proposed and sometimes used in parallel with top-k accuracy. All of these metrics solely evaluate the models in the single-step context, but how the single-step performance translates into the likelihood of success in multi-step planning remains an open question. The evaluation of multi-step planning, on the other hand, tends to have two distinct focuses, either on *efficiency* or on *quality*. Search efficiency has been measured in the success rate of finding pathways with buyable starting materials, as well as average numbers of iterations and node visits. However, as we demonstrate in Figure 1, efficiency metrics like the success rate give little insight into route quality. To evaluate route quality, simple proxies such as route length (Chen et al., 2020; Kishimoto et al., 2019) and average complexity of molecules (Shibukawa et al., 2020) have been used, and so have more complicated heuristics such as tree edit distance to a reference route (Genheden & Bjerrum, 2022).

However, some existing benchmarks (Chen et al., 2020; Genheden & Bjerrum, 2022; Tripp et al., 2022) based on these metrics do not verify if the searched materials can synthesize the target molecule. Although the ideal method for validating the feasibility of starting materials would involve chemical laboratory testing or expert evaluation, these approaches are frequently cost-prohibitive. Consequently, we introduce a complementary matching metric to evaluate retrosynthesis models and search algorithms by comparing predicted starting materials with those retrieved from the dataset during testing. The set-wise exact match of starting materials, as we propose in this work, finds a balance between simplicity and data awareness. It is cheap to compute, easy to implement, and yet provides a good indication of how probable the suggested set will successfully lead to the target. Note that our introduced evaluation metric does not restrict prediction diversity. Although Tripp et al. (2022) propose a metric to evaluate prediction diversity, it does not verify whether the searched starting materials can indeed synthesize the target molecule. As such, a trade-off exists between our metric and theirs. We believe this open problem could stimulate future research to develop new metrics and methods that effectively address both aspects.
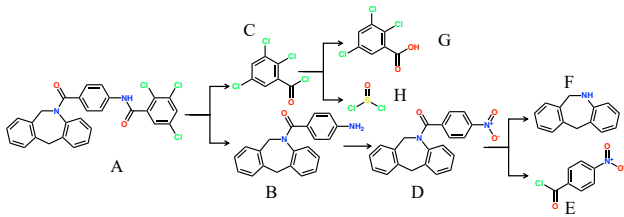
Figure 2. Illustration of *synthetic route*. Given the definition in Eq. (2), the depth of this route is 3, which means the depth of the longest path is 3. A is the desired target molecule to be synthesized. B, C, and D are the intermediates. E, F, G, and H are the starting materials.

3. Background

In this section, we formally define important terminologies used in the rest of the paper, including starting materials and synthetic routes, based on which we define the formulation of retrosynthetic planning.

Starting Material. We denote the space of all chemical molecules as $\mathcal{M}$. Following AiZynthFinder (Genheden et al., 2020), we define the starting materials as a set of commercially purchasable molecules, denoted as $\mathcal{S} \subseteq \mathcal{M}$. ZINC (Sterling & Irwin, 2015) releases the open-source databases of purchasable compounds. We define this list of compounds in these databases as our starting materials.

Synthetic Route. Given the above definitions, a synthetic route can also be organized as a graph-like structure, called reaction graph (Shibukawa et al., 2020; Nguyen & Tsuda, 2021). In the rest of the paper, we use the terminology "reaction graph" instead of "synthetic route". An illustration of a reaction graph (denoted as $\mathcal{G}$) is shown in Figure 2. Here, $\mathcal{G} = \{T, \mathcal{R}, \mathcal{I}, \tau\}$, where $T \in \mathcal{M} \setminus \mathcal{S}$ is the target molecule we desire to synthesize (A in Figure 2), $\mathcal{R} = \{r_1, r_2, \ldots, r_n\} \subseteq \mathcal{S}$ is the set of starting materials (E, F, G, H in Figure 2) that go through a series of reactions τ to synthesize A, and $\mathcal{I} = \{m_1, m_2, \ldots, m_u\} \subseteq \mathcal{M} \setminus \mathcal{S}$ is the set of intermediates (B, C, D in Figure 2) formed from molecules represented by their child nodes, which can react further to produce the molecule represented by their parent nodes. A reaction graph consists of multiple paths from the target molecule to any starting material in the reaction graph. According to the definition, the number of paths is equal to the number of starting materials. We denote paths as l, the set of paths as $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}$, and we have

$$\tau = \tau_{l_1} \cup \tau_{l_2} \cup \cdots \cup \tau_{l_n}, \tag{1}$$

where τ_{l_i} is the set of reactions accompanying path l_i. As illustrated in Figure 2, $A \to B \to D \to E$ is one of the paths in this graph. We denote the depth $\mathcal{D}_{\mathcal{G}}$ of a reaction graph as the length of the longest path in this graph, where

$$\mathcal{D}_{\mathcal{G}} = \max_i \mathcal{D}_{l_i}. \tag{2}$$

The depth of a reaction graph is also the number of steps required to synthesize a molecule from a fixed set of commercially purchasable compounds. Note that in this paper, the default order of the path is in the retrosynthetic (rather than forward) direction.

Single-Step Retrosynthesis. Given a target product molecule $T \in \mathcal{M}$, the goal of one-step retrosynthesis is to predict a set of reactants $\mathcal{R} = \{r_1, r_2, \ldots, r_n\} \subseteq \mathcal{M}$ that can react to synthesize this product, which can be formulated as:

$$T \to \mathcal{R}.$$

Retrosynthetic Planning. Given a target molecule $T \in \mathcal{M}$, the goal of retrosynthetic planning is to search for the starting materials $\mathcal{R} = \{r_1, r_2, \ldots, r_n\} \subseteq \mathcal{S}$ that can synthesize the target molecule through a set of chemical reactions $\tau = \{R_1, R_2, \ldots, R_m\}$, which can be formulated as follows:

$$T \to \mathcal{I} \to \mathcal{R}, \tag{3}$$

where $\mathcal{I} \subseteq \mathcal{M} \setminus \mathcal{S}$ is the set of intermediates.

4. FusionRetro

In this section, we delve into the specifics of our proposed FusionRetro method. We commence by describing how we construct our reaction graphs in Section 4.1, drawing upon the synthetic routes depicted in Figure 2. We then proceed to elaborate on our systematic approach for utilizing informative in-context examples (reactions) from the reaction graph in Section 4.2. This framework involves three principled steps: *encode* molecules into embeddings, *aggregate* the embeddings of molecules through message passing over reaction graphs, and *readout* to predict reactants on current retrosynthesis step. We conclude this section by briefly outlining the practical aspects of our training and inference algorithm in Section 4.3. Figure 3 offers a high-level visual representation of our proposed framework.

4.1. Reaction Graph

In this section, we describe the details of how to construct the reaction graph from the synthetic route.

Task Nodes. First, we introduce the concept of task molecules, which serve as the nodes in our reaction graphs. Specifically, we designate the target molecule T and intermediates $\mathcal{I}$ as task molecules, as these will be expanded during the multi-step planning process. Importantly, because our search process halts when the molecules on the synthetic route are commercially available, these leaf nodes in Figure 2 are not included in our constructed reaction graph.
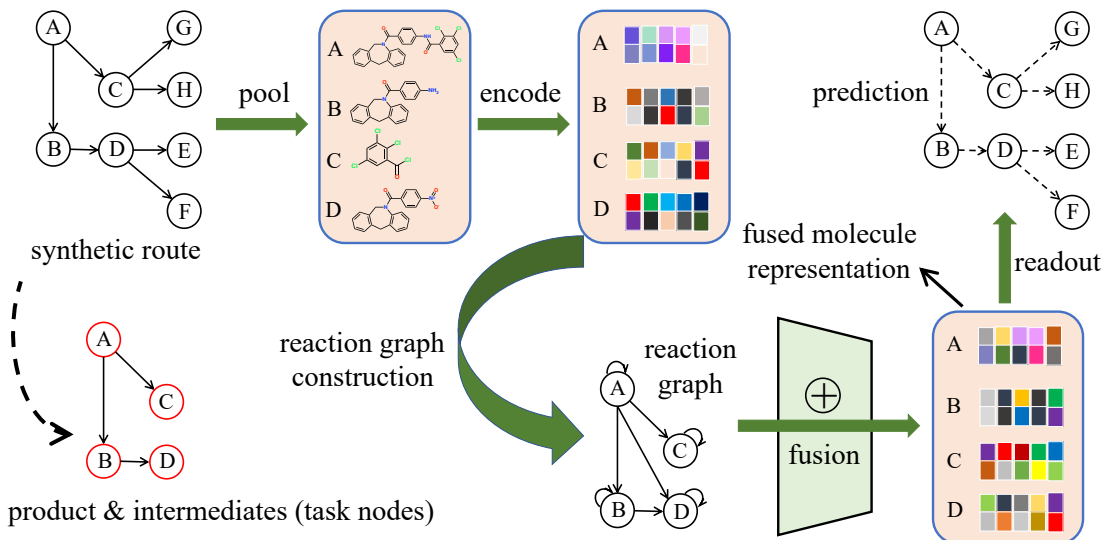
Figure 3. Illustration of our framework. Our framework consists of three modules: *encode*, *aggregation*, and *readout*. The process begins with the construction of a reaction graph from the given synthetic route. After encoding the molecules present in this reaction graph, we utilize the aggregation module to generate the fused molecule representations (FMR). This FMR is used for retrosynthesis prediction.

Graph Construction. In order to explicitly model the contextual information of reactions and intermediates along the synthetic routes, we build reaction graphs among task molecules. We first remove non-task molecules on the leaf nodes in Figure 2. Then, inspired by the dense connection (Huang et al., 2017) between tokens in Transformer (Vaswani et al., 2017), we link each task molecule and its ancestors to construct our reaction graph, which enables us to explicitly model the relational information between task molecules.

4.2. Molecule Representation Fusion

As depicted in Figure 3, a given path consists of several chemical reactions. Inspired by the recent advancements in in-context learning within large language models, we utilize in-context examples—specifically, the reactions preceding the current one—to boost the accuracy of our current prediction. To this end, we propose a well-founded fusion framework. This framework is designed to regulate the information flow and distill representations that seize essential contextual information from the reaction graph.

Another part of our motivation stems from the discrepancy between machine learning methods prevalent in existing works and the actual thought process of chemists. Chemists don't typically think like a search engine – by iteratively applying some rigid one-step expansion with some search criteria. Instead, many of them think in a more holistic way, for example, by taking into account all the intermediate steps when planning the next, for reasons including but not limited to ease of purification. A purely one-step model would likely miss most, if not all, of this contextual information. Thus, in this section, we delve into the specifics of

molecule representation fusion. Particularly, we employ the attention mechanism to generate representations that capture the contextual information of reactions and intermediates along the reaction graph.

Molecule Encoding. Given the reaction graph depicted in Figure 3, the first step involves encoding the molecules in the reaction graph into embeddings using molecule encoders. These encoders can be broadly categorized into sequence-based and graph-based methods. Graph-based models (Shi et al., 2020; Yan et al., 2020; Somnath et al., 2021; Sacha et al., 2021) employ a Message Passing Neural Network (Gilmer et al., 2017) to translate the molecule graph into an embedding vector. On the other hand, sequence-based models (Karpov et al., 2019) leverage the attention mechanism to transform the SMILES representation of the molecule into an embedding matrix. Note that our proposed method is a general framework, and we intentionally omit the details of the encoding process. Instead, we represent the encoder as a function, denoted as ϕ. Therefore, the encoding process can be formulated as follows:

$$\boldsymbol{h}_m = \phi(m), \tag{4}$$

where $m \in \{T\} \cup \mathcal{I}$ and $\boldsymbol{h}_m$ denotes the representation of molecule m.

Representation Fusion. Upon encoding, we carry out a message-passing operation to aggregate the molecule embeddings. This allows us to create fused representations that encapsulate contextual information. Rather than directly employing the weights in the adjacency matrix, we compute the correlation coefficient (Zhang & Zitnik, 2020) to assess the relevance between molecule nodes u and v. Based on
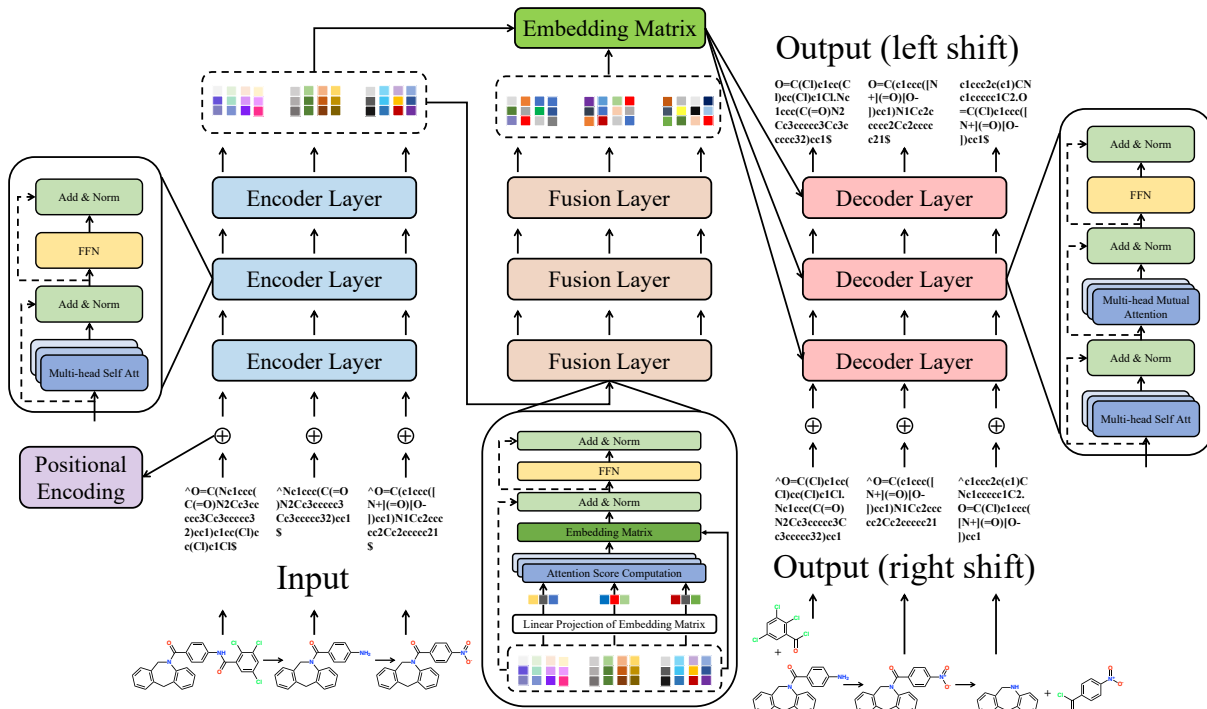
Figure 4. Illustration of our architecture. Our architecture consists of the encoder, decoder, and fusion modules, each of which is composed of several stacked attention layers. In the encoder, we employ self-attention layers to transform the embeddings of input SMILES into latent representations, known as encoder outputs. Subsequently, we utilize the fusion module to attain a fused molecule representation. This fused representation is then fed into the decoder, which yields the final prediction.

these correlation coefficients, we can propagate messages across the weighted reaction graph in a more meaningful manner. To quantify the correlation between two molecule nodes, we make use of an attention mechanism (Veličković et al., 2018; Sukhbaatar et al., 2015; Weston et al., 2015), deriving the coefficients as follows:

$$c(\boldsymbol{h}_u, \boldsymbol{h}_v) = \boldsymbol{h}_u \odot \boldsymbol{h}_v, \qquad (5)$$

where $\odot$ stands for the dot product. In a manner akin to GAT (Veličković et al., 2018), we also normalize the coefficients across all neighbors using the softmax function:

$$\begin{aligned} \alpha(\boldsymbol{h}_u, \boldsymbol{h}_v) &= \text{softmax}_v \left(c(\boldsymbol{h}_u, \boldsymbol{h}_v) \right) \\ &= \frac{\exp \left(c(\boldsymbol{h}_u, \boldsymbol{h}_v) \right)}{\sum_{k \in \mathcal{N}_u} \exp \left(c(\boldsymbol{h}_u, \boldsymbol{h}_k) \right)}, \end{aligned} \qquad (6)$$

where $\mathcal{N}_u$ represents the neighborhood of molecule u in the reaction graph. With this approach, we can quantify the message transmitted along the weighted reaction graph and derive the fused representation as follows:

$$\boldsymbol{h}_u^{'} = \sum_{v \in \mathcal{N}_u} \alpha(\boldsymbol{h}_u, \boldsymbol{h}_v) \boldsymbol{h}_v, \qquad (7)$$

where $\boldsymbol{h}_u^{'}$ denotes the fused molecule representation, which captures the contextual information and thus enables more accurate retrosynthesis predictions in multi-step planning, as will be demonstrated in the experimental section.

Readout. Upon obtaining the fused molecule representation (FMR), we employ both the FMR and the original molecule representation as input to predict the reactants using the decoder. The specifics of the readout process are not discussed here, but we represent it as a function ψ. Thus, the readout process can be expressed as follows:

$$p = \psi(\boldsymbol{h}_u, \boldsymbol{h}_u^{'}), \qquad (8)$$

where p stands for the prediction.

Implementation Details. We implement our proposed module based on Transformer (Karpov et al., 2019) given its use of an end-to-end training paradigm, as depicted in Figure 4. It's important to note that our method is a general framework and can inspire future work to incorporate our framework into other retrosynthesis models.

4.3. Training and Inference

Training. During the training phase, we use the entire reaction graph as input, facilitating parallel computation. Given the SMILES representations of molecules (A, B, C, D) present on the reaction graph, the output should correspond to the SMILES representations of (B+C, D, G+H, E+F). Notably, while predicting B+C, input A is regarded as informative, but the information of (B, D) are not considered. Therefore, during training, the information on child
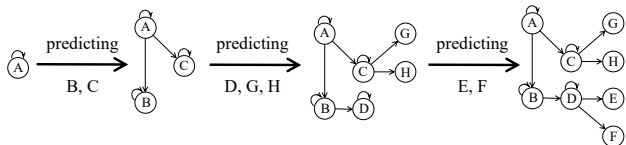
Figure 5. Overview of the inference process. We start from the target molecule A and perform backward chaining to do a series of one-step retrosynthesis predictions until all the final reactants are starting materials.

molecules is excluded when making the current prediction. Thanks to the attention mechanism, our predictions for all reactions along the reaction graph are parallelized during the training phase. This can be achieved by leveraging the adjacency matrix and masking the inputs of child molecules. The loss function can be expressed as follows

$$\mathcal{L}(y, p) = -\sum_{i=1}^{n} \sum_{j=1}^{K} y_{ij} \log(p_{ij}), \qquad (9)$$

where y_{ij} and p_{ij} are the predicted and ground truth values at the j-th position for the i-th target molecule sequence. In other words, the training is parallelized on all the retrosynthesis reactions in the reaction graph.

Inference. During the inference phase, we initiate the process with the target molecule T and apply backward chaining to conduct a series of one-step retrosynthesis predictions until all reactants have been identified as starting materials. After predicting the reactant molecules for each retrosynthesis step, we cross-reference the set of starting materials to verify whether the predicted reactants are indeed starting materials. If they are, we add them to the predicted reactant set. If not, we establish a new path and predict the output for the next step. The inference process concludes once the path set is emptied. This inference process is graphically depicted in Figure 5 and procedurally detailed in Algorithm 1.

5. Experiments

In this section, we evaluate the performance of different retrosynthesis models on our constructed dataset for retrosynthetic planning.

5.1. Dataset Construction

We construct a benchmark for retrosynthetic planning using the public USPTO-full dataset, which consists of 906,164 valid reactions from the original 1,808,937, after removing invalid and duplicate ones. These reactions are used to construct a reaction network (Li & Chen, 2022), treating molecules with an out-degree of zero as target molecules. We use dynamic programming and backtracking to identify

---

Algorithm 1 Inference given a target molecule

1: **Input:** Target molecule T, starting material set $\mathcal{S}$
2: Initialize reactant set $\mathcal{R} = \{\}$, path set $\mathcal{L} = \{\}$
3: Put the initial path $[T]$ into $\mathcal{L}$
4: **while** $\mathcal{L}$ is not a empty set **do**
5:     Take an path l from $\mathcal{L}$
6:     Predict the reactants r_l for expansion given l
7:     **for** reactant $r_l^{(i)}$ in r_l **do**
8:         **if** $r_l^{(i)} \in \mathcal{S}$ **then**
9:             Put $r_l^{(i)}$ into $\mathcal{R}$
10:        **else**
11:            Generate a new path $l' = l + [r_l^{(i)}]$
12:            Put l' into $\mathcal{L}$
13:        **end if**
14:     **end for**
15: **end while**
16: **return** predicted reactant set $\mathcal{R}$

---

all synthetic routes for each target, and following the approach in Chen et al. (2020), we extract the shortest-possible synthetic routes with leaf nodes as starting materials. This process yields synthetic routes for 128,469 molecules. We disregard routes that synthesize target molecules in one step and split the remaining molecules into training, validation, and test datasets in an 80%/10%/10% ratio. This results in 46,458 samples for training, 5,803 for validation, and 5,838 for testing. Note that the target molecules in the training set, validation set, and test set do not intersect. We call our benchmark RetroBench. Detailed statistics of the dataset can be found in Appendix A.

5.2. Experiment Setup

Evaluation Protocol. As previously discussed, current search algorithms (Segler et al., 2018; Kishimoto et al., 2019; Chen et al., 2020; Kim et al., 2021; Xie et al., 2022; Yu et al., 2022) primarily utilize search success rate as their evaluation metric, without verifying if the identified starting materials can indeed synthesize the target molecule. In this study, we propose a new evaluation metric: the set-wise exact match between the proposed starting materials and the ground truth. For a given target molecule, we carry out a series of one-step retrosynthesis predictions and employ search algorithms to select the most promising reactant candidates for expansion, until all leaf nodes have been identified as starting materials. We use the starting materials sourced from our constructed reaction network as the ground truth and compare them to the starting materials identified through the search. The match is based on a basic comparison of the InChiKey of the molecule, as used by AiZynthFinder (Genheden et al., 2020). It's important to note that a particular target molecule may have multiple synthetic routes in the test set. We consider it an accurate

Table 1. Summary of retrosynthetic planning results in terms of exact match accuracy (%).

Search Algorithm	Retro*					Retro*-0					Greedy DFS
	Top-1	Top-2	Top-3	Top-4	Top-5	Top-1	Top-2	Top-3	Top-4	Top-5	Top-1
Template-based											
Retrosim (Coley et al., 2017)	35.1	40.5	42.9	44.0	44.6	35.0	40.5	43.0	44.1	44.6	31.5
Neuralsym (Segler & Waller, 2017)	**41.7**	**49.2**	52.1	53.6	54.4	**42.0**	**49.3**	52.0	53.6	54.3	**39.2**
GLN (Dai et al., 2019)	39.6	48.9	**52.7**	**54.6**	**55.7**	39.5	48.7	**52.6**	**54.5**	**55.6**	38.0
Template-free											
G2Gs (Shi et al., 2020)	5.4	8.3	9.9	10.9	11.7	4.2	6.5	7.6	8.3	8.9	3.8
GraphRetro (Somnath et al., 2021)	15.3	19.5	21.0	21.9	22.4	15.3	19.5	21.0	21.9	22.2	14.4
Megan (Sacha et al., 2021)	18.8	29.7	37.2	42.6	45.9	19.5	28.0	33.2	36.4	38.5	32.9
Transformer (Karpov et al., 2019)	31.3	40.4	44.7	47.2	48.9	31.2	40.5	45.1	47.3	48.7	26.7
FusionRetro	**37.5**	**45.0**	**48.2**	**50.0**	**50.9**	**37.5**	**45.0**	**48.3**	**50.2**	**51.2**	**33.8**

match when the predicted starting material set aligns with at least one of the multiple ground truths. Additionally, we implement a pruning search, halting the search when the length of the predicted synthetic route surpasses the depth of the ground truth synthetic route. Utilizing our evaluation metric allows us to compare the performances of different retrosynthesis models in conjunction with various search algorithms, thereby providing a benchmark for future studies.

Setting and Baselines. We evaluate the effectiveness of our proposed retrosynthesis method in conjunction with three different search algorithms for retrosynthetic planning. This approach is benchmarked against existing single-step retrosynthesis models, which can be broadly categorized into two groups: template-based and template-free models. Each model is trained using the reactions in our training dataset. Upon completion of the retrosynthesis training, we employ the Retro* (Chen et al., 2020), Retro*-0, and Greedy DFS search algorithms. For all baselines, except for Transformer, we adhere to their original experimental setups, including hyperparameters and data processing, as described in their respective papers. These experiments are conducted using their publicly available codes. Transformer is implemented using Pytorch (Paszke et al., 2019), and we re-tuned the learning rate due to the spike phenomenon observed with the learning rate reported in the original paper. The template-based baseline approaches we consider include Retrosim (Coley et al., 2017), Neuralsym (Segler & Waller, 2017), and GLN (Dai et al., 2019). We also evaluate end-to-end template-free approaches such as Transformer (Karpov et al., 2019) and Megan (Sacha et al., 2021), as well as semi-template-based models like G2Gs (Shi et al., 2020) and GraphRetro (Somnath et al., 2021). Our framework is depicted in Figure 4. For all hyperparameters, except for the learning rate (due to the spike phenomenon), we adhere to the settings reported in the publicly released Transformer

code and do not perform any additional hyperparameter tuning. Detailed information on the hyperparameters can be found in Appendix B.1. Our proposed model, FusionRetro, is trained using 2 NVIDIA Tesla V100 GPUs.

5.3. Results

Comparison with Template-free Baselines. The primary results are presented in Table 1. It's clear that our proposed model, FusionRetro, outperforms other template-free baseline methods. Further insights can be drawn from Figure 6, which shows that as the depth of the ground truth synthetic routes increases, the performance gap between the Transformer and FusionRetro generally widens. This demonstrates the value of incorporating context information for representation fusion. In essence, these results indicate that our proposed model consistently performs better than Transformer, particularly in predicting long synthetic routes.

Analysis of the Benchmark. The performance of baseline models on our benchmark does not align well with single-step retrosynthesis predictions on the USPTO-50K dataset. Current two-stage semi-template-based models (Shi et al., 2020; Somnath et al., 2021) either outperform or match template-based and template-free models on USPTO-50K single-step retrosynthesis prediction, yet perform poorly on our benchmark. One main factor is that approximately 95% of reactions in the USPTO-50K dataset have only one reaction center due to heavy filtering, whereas in our constructed dataset, around 30% of reactions have multiple reaction centers. Upon examining the open-source code of G2Gs, we found that it can only handle cases with one reaction center, which explains its weak performance on our benchmark. The performance of template-free models is not impacted by the number of reaction centers. Additionally, we present the results of single-step retrosynthesis
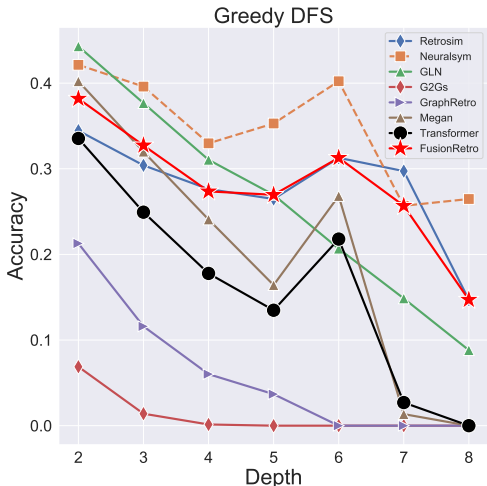
Figure 6. The test accuracy of retrosynthesis models combined with Greedy DFS at different depths of the ground truth synthetic routes. *Red stars* (★) denotes our method (FusionRetro) and *Black circles* (●) represents Transformer.

Table 2. Summary of retrosynthesis prediction results in terms of exact match accuracy (%).

Methods	Top-k accuracy %			
	1	3	5	10
G2Gs	16.5	27.8	33.1	40.4
GraphRetro	48.3	58.4	60.5	62.4
Transformer	55.8	70.3	74.8	78.9
Retrosim	56.5	65.8	69.0	73.1
Megan	59.5	73.9	77.9	81.7
Neuralsym	**63.0**	73.3	76.0	78.6
GLN	62.9	**74.1**	**78.4**	**82.7**

predictions on our constructed test dataset in Table 2. These results align with those of retrosynthetic planning in Table 1, leading us to conclude that single-step accuracy plays a crucial role in multi-step planning as well.

Analysis for the Depth of Routes. As illustrated in Figure 6, the accuracy of prediction tends to decrease as the depth of synthetic routes increases. However, our model exhibits a slower rate of performance degradation compared to other baseline models. This indicates the strength of our approach, which uses contextual information for representation fusion, particularly when predicting long synthetic routes.

5.4. Case Study

Figure 7 provides a visual comparison of predictions made by FusionRetro and Transformer. The upper section of the figure displays accurate predictions made by FusionRetro,
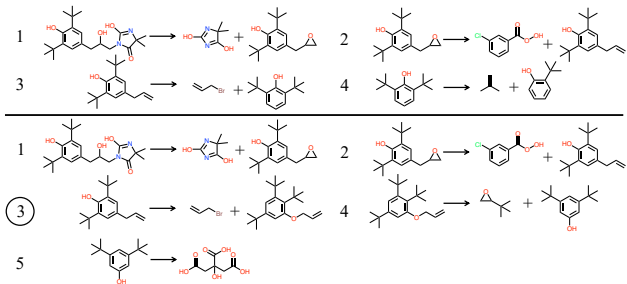

Figure 7. We split the predicted synthetic route into individual reactions. The correct synthetic route predicted by FusionRetro is depicted at the top, while the route predicted by the Transformer model is displayed at the bottom.

while the lower section shows incorrect predictions made by Transformer. It is evident from the figure that Transformer inaccurately predicts the third-step retrosynthesis reaction. Although a search can still identify starting materials, these materials may not be capable of synthesizing the target molecule. Therefore, the performance of retrosynthesis prediction is crucial for effective retrosynthetic planning.

6. Conclusion and Future Work

In this paper, we propose FusionRetro, a novel framework for retrosynthetic planning that exploits crucial context information on the synthetic route by principled representation fusion. FusionRetro is the first method in this field that takes context information into account, greatly boosting the performance for realistic multi-step planning. We further introduce new benchmarks for better evaluation of retrosynthesis models, especially for practical multi-step planning settings. Extensive experiments demonstrate FusionRetro can consistently achieve significantly superior performance across several measurements. We hope our approach can shed light on the research of data-driven retrosynthetic planning, and inspire more studies toward the practical multi-step scenario. Besides, our approach can be viewed as in-context learning and can inspire more works to further explore in-context learning techniques in large language models for multi-step planning. In this way, we can enrich the decision-making process with valuable context-driven inputs.

Acknowledgements

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

Chen, B., Shen, T., Jaakkola, T. S., and Barzilay, R. Learning to make generalizable and diverse predictions for retrosynthesis. *arXiv preprint arXiv:1910.09688*, 2019.

Chen, B., Li, C., Dai, H., and Song, L. Retro*: learning retrosynthetic planning with neural guided a* search. In *International Conference on Machine Learning*, 2020.

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

Coley, C. W., Rogers, L., Green, W. H., and Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Central Science*, 3(12):1237–1245, 2017.

Coley, C. W., Green, W. H., and Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of Chemical Research*, 51(5):1281–1289, 2018.

Dai, H., Li, C., Coley, C., Dai, B., and Song, L. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, 2019.

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Genheden, S. and Bjerrum, E. Paroutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022.

Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist, O., and Bjerrum, E. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1):1–9, 2020.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

Han, P., Zhao, P., Lu, C., Huang, J., Wu, J., Shang, S., Yao, B., and Zhang, X. Gnn-retro: Retrosynthetic planning with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Hong, S., Zhuo, H. H., Jin, K., and Zhou, Z. Retrosynthetic planning with experience-guided monte carlo tree search. *arXiv preprint arXiv:2112.06028*, 2021.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.

Karpov, P., Godin, G., and Tetko, I. V. A transformer model for retrosynthesis. In *International Conference on Artificial Neural Networks*, 2019.

Kim, J., Ahn, S., Lee, H., and Shin, J. Self-improved retrosynthetic planning. In *International Conference on Machine Learning*, 2021.

Kipf, T. N. and Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Kishimoto, A., Buesser, B., Chen, B., and Botea, A. Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In *Advances in Neural Information Processing Systems*, 2019.

Li, B. and Chen, H. Prediction of compound synthesis accessibility based on reaction knowledge graph. *Molecules*, 27(3):1039, 2022.

Lin, K., Xu, Y., Pei, J., and Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.*, 11(12):3355–3364, March 2020. ISSN 2041-6539. doi: 10.1039/C9SC03666K. URL https://pubs.rsc.org/en/content/articlelanding/2020/sc/c9sc03666k. Publisher: The Royal Society of Chemistry.

Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P., and Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Central Science*, 3 (10):1103–1113, 2017.

Meng, Z., Zhao, P., Yu, Y., and King, I. A unified view of deep learning for reaction and retrosynthesis prediction: Current status and future challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2023.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Nguyen, D. H. and Tsuda, K. A generative model for molecule generation based on chemical reaction trees. *arXiv preprint arXiv:2106.03394*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Sacha, M., Błaz, M., Byrski, P., Dabrowski-Tumanski, P., Chrominski, M., Loska, R., Włodarczyk-Pruszynski, P., and Jastrzebski, S. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284, 2021.

Schwaller, P., Petraglia, R., Zullo, V., Nair, V. H., Haeuselmann, R. A., Pisoni, R., Bekas, C., Iuliano, A., and Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.*, 11:3316–3325, 2020. doi: 10. 1039/C9SC05704H. URL http://dx.doi.org/10.1039/C9SC05704H.

Segler, M. H. and Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

Segler, M. H., Preuss, M., and Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604, 2018.

Seidl, P., Renz, P., Dyubankova, N., Neves, P., Verhoeven, J., Wegner, J. K., Hochreiter, S., and Klambauer, G. Modern hopfield networks for few-and zero-shot reaction prediction. *arXiv preprint arXiv:2104.03279*, 2021.

Shi, C., Xu, M., Guo, H., Zhang, M., and Tang, J. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning*, 2020.

Shibukawa, R., Ishida, S., Yoshizoe, K., Wasa, K., Takasu, K., Okuno, Y., Terayama, K., and Tsuda, K. Compret: a comprehensive recommendation framework for chemical synthesis planning with algorithmic enumeration. *Journal of Cheminformatics*, 12(1):1–14, 2020.

Somnath, V. R., Bunne, C., Coley, C., Krause, A., and Barzilay, R. Learning graph models for retrosynthesis prediction. In *Advances in Neural Information Processing Systems*, 2021.

Sterling, T. and Irwin, J. J. Zinc 15–ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, 2015.

Sukhbaatar, S., Weston, J., Fergus, R., et al. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, 2015.

Sun, R., Dai, H., Li, L., Kearnes, S., and Dai, B. Towards understanding retrosynthesis by energy-based models. In *Advances in Neural Information Processing Systems*, 2021.

Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11, 2020.

Tripp, A., Maziarz, K., Lewis, S., Liu, G., and Segler, M. Re-evaluating chemical synthesis planning algorithms. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Weston, J., Chopra, S., and Bordes, A. Memory networks. In *International Conference on Learning Representations*, 2015.

Xie, S., Yan, R., Han, P., Xia, Y., Wu, L., Guo, C., Yang, B., and Qin, T. Retrograph: Retrosynthetic planning with graph search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

Yan, C., Ding, Q., Zhao, P., Zheng, S., Yang, J., Yu, Y., and Huang, J. Retroxpert: Decompose retrosynthesis prediction like a chemist. In *Advances in Neural Information Processing Systems*, 2020.

Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. Do transformers really perform badly for graph representation? In *Advances in Neural Information Processing Systems*, 2021.

Yu, Y., Wei, Y., Kuang, K., Huang, Z., Yao, H., and Wu, F. Grasp: Navigating retrosynthetic planning with goal-driven policy. In *Advances in Neural Information Processing Systems*, 2022.

Zhang, X. and Zitnik, M. Gnnguard: Defending graph neural networks against adversarial attacks. In *Advances in Neural Information Processing Systems*, 2020.

Zheng, S., Rao, J., Zhang, Z., Xu, J., and Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, 60(1):47–55, 2019.

A. Datasets Details

Table 3. The number of target molecules in training/validation/test datasets in term of the shortest depths to synthesize the target molecules.

#Molecules ╲ Depth Dataset	2	3	4	5	6	7	8	9	10	11	12	13
Training	22,903	12,004	5,849	3,268	1,432	594	276	107	25	0	0	0
Validation	2,862	1,500	731	408	179	74	34	13	2	0	0	0
Test	2,862	1,500	731	408	179	74	34	13	2	32	2	1

B. Reproducibility

B.1. Implementation Details

We use Pytorch (Paszke et al., 2019) to implement FusionRetro. The codes of baselines are implemented referring to the implementation of Retrosim[1], Neuralsym[2], GLN[3], G2Gs[4], GraphRetro[5], Transformer[6], and Megan[7]. All the experiments of baselines are conducted on a single NVIDIA Tesla V100 with 32GB memory size. The software that we use for experiments are Python 3.6.8, pytorch 1.9.0, pytorch-scatter 2.0.9, pytorch-sparse 0.6.12, numpy 1.19.2, torchvision 0.10.0, CUDA 10.2.89, CUDNN 7.6.5, einops 0.4.1, and torchdrug 0.1.3.

B.2. Hyperparameter Details

Table 4. The hyper-parameters for FusionRetro.

max length	200
embedding size	64
encoder layers	3
decoder layers	3
fusion layers	3
attention heads	10
FFN hidden	512
dropout	0.1
epochs	4000
batch size	64
warmup	16000
lr factor	20

C. More results

---

[1] https://github.com/connorcoley/retrosim

[2] https://github.com/linminhtoo/neuralsym

[3] https://github.com/Hanjun-Dai/GLN

[4] https://torchdrug.ai/docs/tutorials/retrosynthesis

[5] https://github.com/vsomnath/graphretro

[6] https://github.com/bigchem/synthesis
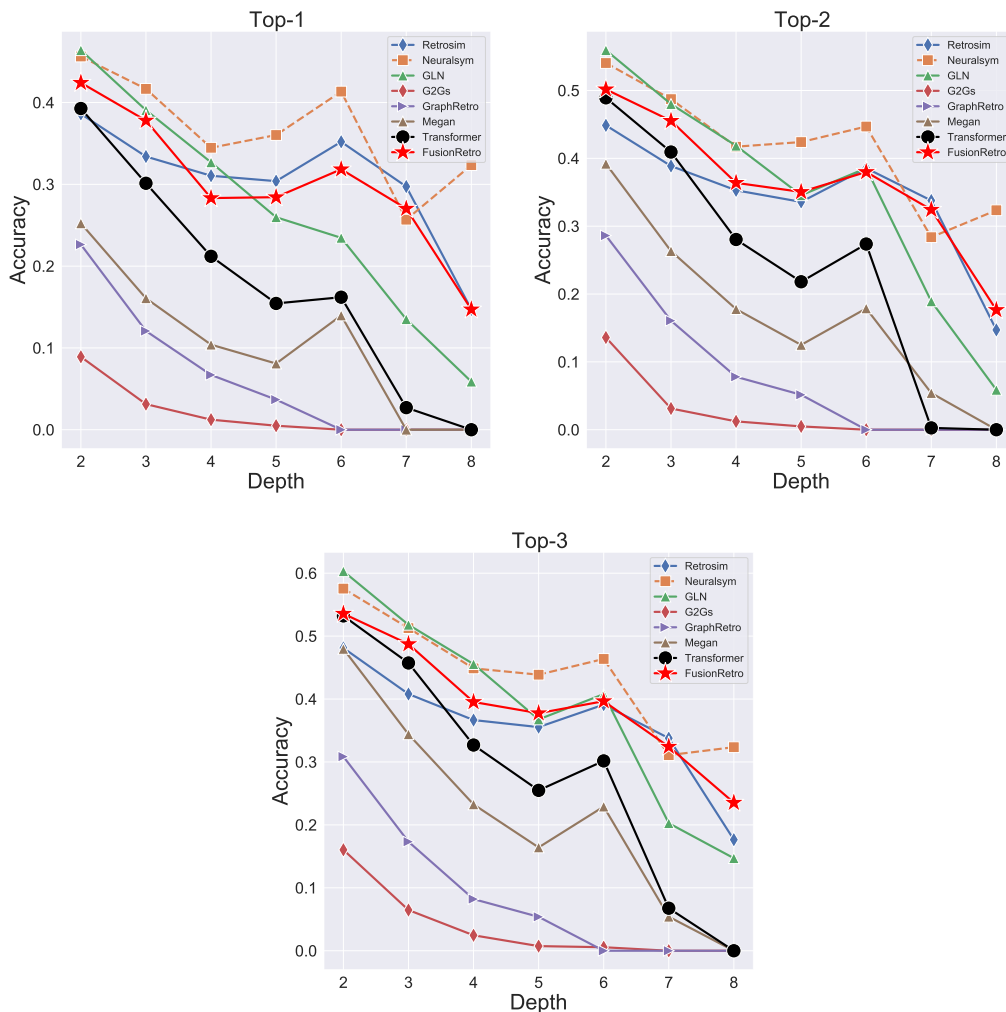
[7] https://github.com/molecule-one/megan

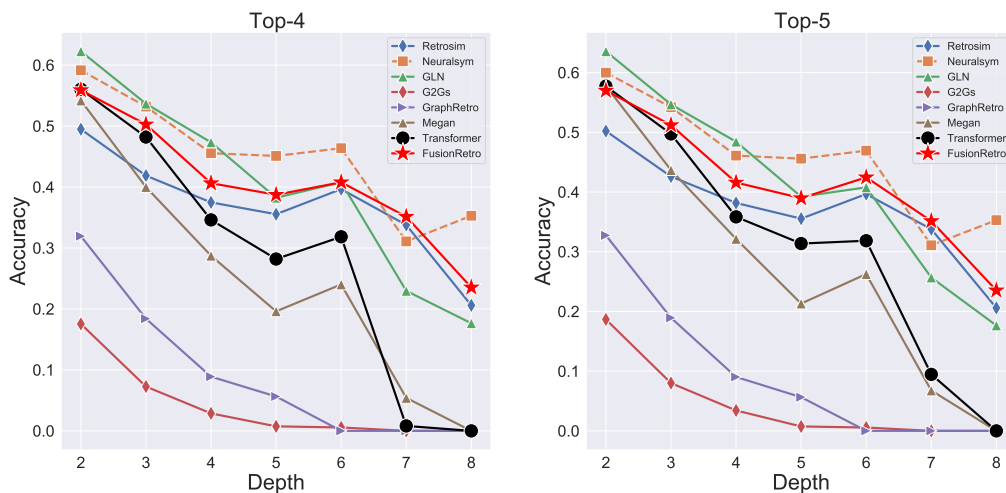*Figure 8.* The top-1, top-2, and top-3 test accuracy in terms of depth.

Figure 9. The top-4 and top-5 test accuracy in terms of depth.