Advancing Retrosynthesis with Retrieval-Augmented Graph Generation

Anjie Qiao[*1], **Zhen Wang**[*†1,2], **Jiahua Rao**[1], **Yuedong Yang**[†1], **Zhewei Wei**[3]

[1]Sun Yat-sen University, Guangzhou, Guangdong, China
[2]Guangdong Province Key Laboratory of Computational Science
[3]Renmin University of China, Beijing, China
qiaoanj@mail2.sysu.edu.cn, {wangzh665,raojh7,yangyd25}@mail.sysu.edu.cn, zhewei@ruc.edu.cn

Abstract

Diffusion-based molecular graph generative models have achieved significant success in template-free, single-step retrosynthesis prediction. However, these models typically generate reactants from scratch, often overlooking the fact that the scaffold of a product molecule typically remains unchanged during chemical reactions. To leverage this useful observation, we introduce a retrieval-augmented molecular graph generation framework. Our framework comprises three key components: a retrieval component that identifies similar molecules for the given product, an integration component that learns valuable clues from these molecules about which part of the product should remain unchanged, and a base generative model that is prompted by these clues to generate the corresponding reactants. We explore various design choices for critical and under-explored aspects of this framework and instantiate it as the **R**etrieval-**A**ugmented **R**etro**B**ridge (RARB). RARB demonstrates state-of-the-art performance on standard benchmarks, achieving a 14.8% relative improvement in top-1 accuracy over its base generative model, highlighting the effectiveness of retrieval augmentation. Additionally, RARB excels in handling out-of-distribution molecules, and its advantages remain significant even with smaller models or fewer denoising steps. These strengths make RARB highly valuable for real-world retrosynthesis applications, where extrapolation to novel molecules and high-throughput prediction are essential.

1 Introduction

In recent years, machine learning methods have shown great promise in accelerating the *de novo* drug discovery process (Blakemore et al. 2018). A crucial stage of this process is retrosynthesis prediction—designing the synthesis routes that can lead to the target molecules through a series of chemical reactions (Strieth-Kalthoff et al. 2020). In this paper, we focus on template-free, single-step prediction, namely identifying direct precursor(s) of a given product molecule, which can serve as a building block in designing the entire synthesis pathway (Segler, Preuss, and Waller 2018; Zhong et al. 2023).

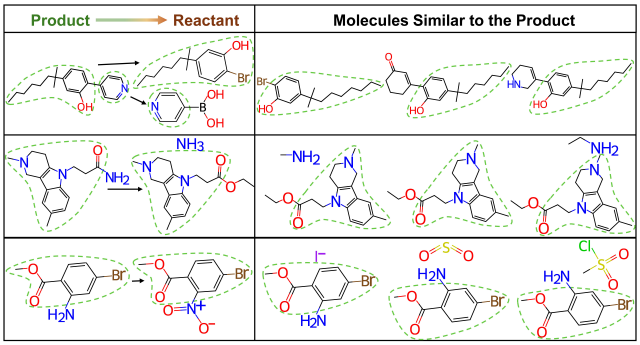

Figure 1: Randomly sampled chemical reactions and retrieved molecules similar to the given product, showcasing the intuition that they can provide clues about which parts of the product should remain unchanged.

Many existing methods formulate this prediction task as a sequence-to-sequence translation problem based on the SMILES (Weininger 1988) of reactants and products (Zheng et al. 2019; Tetko et al. 2020; Kim et al. 2021; Wan et al. 2022; Zhong et al. 2022; Zeng et al. 2024; Han et al. 2024). Despite their advantages, such as simplicity and efficiency, representing molecules as sequences can lose important structural information from the molecular graph, which might lead to less accurate predictions (Shi et al. 2020).

However, turning to graph-to-graph translation problem is not a free lunch since graph generation is challenging, mainly due to the combinatorial nature of graph data that often results in a high intrinsic dimension. As even a single error in predicting a bond or atom type can make the whole synthesized graph incorrect, traditional one-shot graph generation methods struggle to achieve satisfactory performance (Igashov et al. 2024). Recently, diffusion models have become the state-of-the-art for molecular graph generation, where an iterative process decomposes the original one-shot generation into a series of small graph edits (Vignac et al. 2023). Therefore, some recent works utilize diffusion models to frame retrosynthesis prediction as a conditional graph generation problem, achieving unprecedented successes (Wang et al. 2023; Laabid et al. 2024; Igashov et al. 2024).

Although the decomposition strategy of diffusion models

[*]These authors contributed equally.
[†]Co-corresponding authors.

effectively addresses the high-dimensional nature of graph data to some extent, we note that these methods approach retrosynthesis prediction by generating molecular graphs from scratch, thereby overlooking a common phenomenon in chemical reaction: the product's scaffold largely remains unchanged during the reaction, with changes occurring primarily at the reaction center (Fang et al. 2023). Consequently, what needs to be conditionally generated are the small subgraphs that change in the chemical reaction, if given both the product itself and the unchanged part of it. This observation motivates us to introduce an idea (visualized in Fig. 1): We augment the conditioned product with similar molecules, where analyzing their substructures or fragments can provide valuable clues about what should remain unchanged.

Therefore, we propose a retrieval-augmented molecular graph generation framework specially designed for template-free, single-step retrosynthesis prediction task, as illustrated in Fig. 2. Our framework consists of three main components: a retrieval component, an integration component, and a base generative model. Given a product, the retrieval component uses it as the query to search for similar molecules from an external dataset of commercially available molecules. The retrieval results are processed by the integration component to extract useful clues, which are then fed into the base generative model. With these clues, the generative model is expected to achieve more competitive performance.

Nevertheless, applying Retrieval-Augmented Generation (RAG) to retrosynthesis prediction is nontrivial. For our retrieval component, determining which external dataset to use and how to effectively retrieve molecular graphs requires exploration, given that the chemical space is larger and more diverse than natural images, and there is no direct counterpart of CLIP encoders (Radford et al. 2021) for graph data. For our integration component, it is essential to study how to represent the retrieved molecules and which neural architecture is suitable for the augmented generative model. The cross-attention mechanisms that successfully serve text-to-image generation (Peebles and Xie 2023) may not directly translate for facilitating interactions among tokens (i.e., atoms) of different graphs.

Hence, we explore various potential design choices through pilot experiments and comprehensive analysis to instantiate our framework as **R**etrieval-**A**ugmented **R**etro**B**ridge (RARB), using the Markov bridge-based RetroBridge (Igashov et al. 2024) as the base generative model. We compare RARB with the state-of-the-art retrosynthesis prediction methods on standard benchmarks. RARB not only achieves the best performance in regular settings but also show that retrieval augmentation can provide advantages for handling out-of-distribution molecules. Moreover, RARB maintains remarkable advantages even with smaller models or fewer denoising steps. Meanwhile, our introduced retrieval augmentation preserves both the diversity and efficiency of the adopted base generative model at nearly the same level.

We summarize our contributions as follows:

- We propose a retrieval-augmented molecular graph generation framework for retrosynthesis prediction that can be easily plugged into various generative models, seamlessly enhancing their performance.
- We thoroughly investigate the design choices for our proposed framework and gain useful insights to develop RARB as a practical instance.
- RARB achieves state-of-the-art performance, with a notable 14.8% relative improvement in top-1 accuracy over the base generative model. Additionally, RARB shows advantages in handling out-of-distribution products.

2 Related Work

Retrosynthesis Modeling. Recent retrosynthesis prediction methods can be categorized into three main groups based on their dependency on prior chemical knowledge (Zhong et al. 2023; Liu et al. 2023). (1) *Template-based* methods formulate retrosynthesis as a classification problem, selecting a proper reaction template from a predefined set of candidates, which significantly limits their generalization ability (Zhong et al. 2023). (2) *Semi-template* methods decompose retrosynthesis into two stages: first, identifying the reaction center to obtain intermediate molecules known as synthons, and second, converting these synthons to reactants. However, these methods are constrained by their definition of the reaction center and are prone to error propagation (Zhong, Yang, and Chen 2023). (3) *Template-free*, fully end-to-end methods are the most scalable, as they do not rely on prior chemical knowledge and generate the target reactants directly. These methods typically use SMILES (Zhang et al. 2024) or molecular graph for data representation, leading to sequence-base methods (Zheng et al. 2019; Tetko et al. 2020; Kim et al. 2021; Zhang et al. 2024) and graph-based methods (Igashov et al. 2024; Laabid et al. 2024), respectively. In this work, we propose RARB, an end-to-end template-free graph-based method. Despite being template-free, its retrieval component can provide hints similar to semi-template methods.

Generative Diffusion Models. Generative diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020), which are parameterized Markov chain trained to reverse a pre-defined forward diffusion process that gradually corrupts training data into noise, have shown promising results across various domains (Kong et al. 2021; Dhariwal and Nichol 2021; Vignac et al. 2023; Ho et al. 2022; Li et al. 2022). Recently, researchers have extended these models to handle discrete random variables (Austin et al. 2021), particularly molecular graphs (Vignac et al. 2023; Hoogeboom et al. 2022). Discrete diffusion model have been applied to retrosynthesis, treating chemical reactions as a conditional graph generation task (Wang et al. 2023; Laabid et al. 2024). Igashov et al. (2024) proposed a Markov bridge-based approach to model the probabilistic dependency between the spaces of products and reactants, outperforming vanilla diffusion models for conditional graph generation. However, existing diffusion model-based methods often overlook the observation that the molecular graph remains largely unchanged during chemical reactions (Zeng et al. 2024). Our proposed RARB introduces a retrieval-

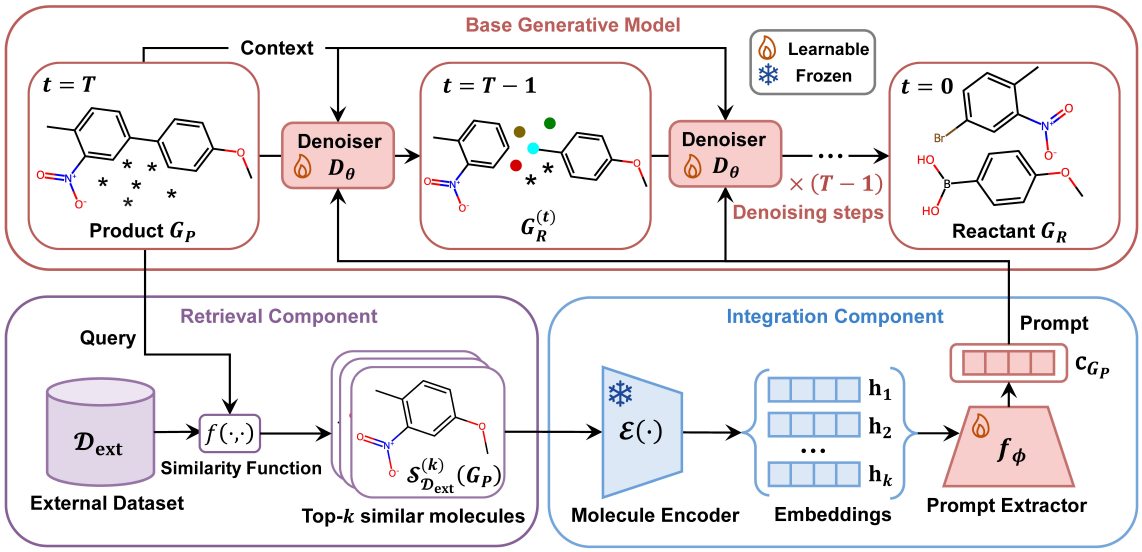

Figure 2: Overview of our retrieval-augmented molecular graph generation framework. The '*' symbol indicates "dummy" atoms, which represent atoms that may be denoised into real atoms in the final reactant.

augmented strategy to leverage common substructures in similar molecules, framing retrosynthesis as a substructure-level generative task.

Retrieval-Augmented Generation. Despite the impressive performance of recently proposed generative models, they still face challenges such as outdated knowledge, lack of long-tailed knowledge, and significant demands in energy consumption and training time (Zhao et al. 2024). Retrieval-Augmented Generation (RAG) aims to mitigate these issues using external memory and flexible retrieval dataset (Carlini et al. 2021; Mallen et al. 2023; Kang et al. 2024). Most existing works focus on text-related tasks facilitated by large language models, with limited exploration into other modalities (Xu et al. 2021). Blattmann et al. (2022) proposed retrieval-augmented diffusion model for image synthesis task. To the best of our knowledge, our proposed RARB is the first successful attempt to apply a retrieval-augmented diffusion model to reactants generation.

3 Methodology

In our work, we represent a molecular graph as $G = (\mathbf{X}, \mathbf{E})$, where $\mathbf{X} \in \mathbb{R}^{N \times K_a}$ is the matrix of atom feature, and $\mathbf{E} \in \mathbb{R}^{N \times N \times K_b}$ is the matrix of bond feature. Here, N denotes the number of atoms, and each atom and bond have K_a and K_b categories, respectively. Thus, we can denote each given product as a graph G_{P} and, without loss of generality, represent a set of its reactant molecules as G_{R}, where each molecule corresponds to a connected component. Given that a product can have multiple sets of reactants that lead to its formation, we frame the template-free, single-step retrosynthesis prediction task as estimating the conditional density function $p(G_{\text{R}}|G_{\text{P}})$[1] based on a finite sample of observed graph pairs, denoted as $\mathcal{D}_{\text{obs}} = \{(G_{\text{P}_i}, G_{\text{R}_i})\}_{i=1}^{|\mathcal{D}_{\text{obs}}|}$.

To this end, we propose a retrieval-augmented molecular graph generation framework. As depicted in Fig. 2, our framework includes a retrieval component, an integration component, and a base generative model that work synergistically to enable the sampling of G_{R} for a given G_{P}. A key feature is the use of an external dataset of commercially available molecules, denoted as $\mathcal{D}_{\text{ext}} = \{G_j\}_{j=1}^{|\mathcal{D}_{\text{ext}}|}$. Intuitively, the retrieval component searches for "relevant" graphs from $\mathcal{D}_{\text{ext}}$ for each given G_{P}, while the integration component extracts helpful information from the search results to prompt the base generative model. Any generative modeling method that can incorporate this additional information can serve as our base model, with the prompt seamlessly integrated to enhance its generative capabilities. Below, we elaborate on these three components, focusing on critical design choices and how they culminate in the development of our specific instance, RARB.

3.1 Retrieval Component

Given a query, i.e., a product G_{P}, this component searches $\mathcal{D}_{\text{ext}}$ and returns a specified number of molecules. We denote the search results for k molecules as $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$. Determining both the appropriate external dataset $\mathcal{D}_{\text{ext}}$ and the strategy for producing $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$ is crucial.

External Dataset. Based on our observation in Fig. 1, the quality of retrieved molecules directly influences the substructure information extracted as clues about which part of G_{P} should remain unchanged. Fang et al. (2023) also attempt to extract common substructures from an external dataset, allowing them to focus on predicting the SMILES of the remaining fragments. Their external dataset is derived from their product-reactant pairs that are used for model training. However, recent RAG research in CV and NLP

[1]For simplicity, we use this graph notation to denote both the random variable and its realization.

has shown that relying solely on the training data for retrieval can potentially limit the model's generalization capacity (Blattmann et al. 2022). Hence, we conjecture that an external dataset $\mathcal{D}_{\text{ext}}$ that includes a diverse and extensive collection of molecules is more likely to yield common substructures resembling the product's scaffold.

To verify our conjecture, we train and evaluate RARB on the USPTO-50k dataset. Accordingly, we compare two distinct choices for $\mathcal{D}_{\text{ext}}$: **(i)** USPTO-50k: A subset of all reactants in the training split is used as $\mathcal{D}_{\text{ext}}$ for model training and validation, and a subset of all reactant in both the training and validation splits is used as $\mathcal{D}_{\text{ext}}$ for testing. **(ii)** USPTO-applications: a subset of all reactants collected from the USPTO 2001-2016 applications, which is significantly larger and more diverse than USPTO-50k.

Pilot Experiment 1: We compare the performance of RARB *using these two choices of $\mathcal{D}_{ext}$. USPTO-applications dataset demonstrates improved performance across all metrics, with the top-1 accuracy increasing by 16.6% compared to that using USPTO-50k.*

More details can be found in Appendix C.1.

Retrieval Strategy. Conventionally, $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$ can be defined as the top-k instances in $\mathcal{D}_{\text{ext}}$ that maximizes a specific similarity function w.r.t. the query, i.e., $f(G_{\text{P}}, \cdot)$. When calculating similarity, we need to design how to represent the molecule and select an appropriate similarity metric. There are two main strategies: **(i)** computing the cosine similarity between two vector representations encoded by a learnable encoder, and **(ii)** computing a general or domain-specific similarity metric based on manually designed features.

Most RAG methods in the CV and NLP domains adopt the former strategy. However, despite recent advances in representation learning for molecular graphs (Zhou et al. 2023; Qiang et al. 2023), there is no universally adopted "standard" molecule encoder akin to CLIP in those domains. This raises a concern: will the representation issue become an obstacle to applying RAG to molecular graphs?

We argue that features based on structural patterns predefined by domain experts can meet the need since we aim to retrieve molecules that share similar sub-structures, particularly the scaffold, with the query. As a potential design, we consider using *Morgan* fingerprints to obtain the bit vector $\mathbf{v}_j$ for molecule G_j and use the *Tanimoto* coefficient (Bajusz, Rácz, and Héberger 2015) as similarity function $f(G_i, G_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\|^2 + \|\mathbf{v}_j\|^2 - \mathbf{v}_i \cdot \mathbf{v}_j}$. We conduct a preliminary experiment to evaluate this design:

Pilot Experiment 2: We compute $f(G_P, G_j), G_j \in \{G_R\} \cup \mathcal{D}_{ext}$ and rank all G_js in descending order. Then, we evaluate the ranking of G_R. Our design, namely, Morgan Fingerprint (radius: 2, bit: 4096) results in an averaged ranking 1.22, which implies the validity of our design.

In this pilot experiment, the rationale is that, if we want the molecules retrieved based on G_{P} to be informative about the ground-truth reactants G_{R}, then $f(G_{\text{P}}, G_{\text{R}})$ should be higher than for most of the candidates. More details about this pilot experiments can be found in Appendix C.2. It is worth noticing that our RAG framework neither requires nor expects $\mathcal{D}_{\text{ext}}$ to include the ground-truth reactant molecules. More importantly, in Sec. 4.1, we explicitly exclude the ground-truth reactants from our adopted $\mathcal{D}_{\text{ext}}$.

3.2 Integration Component

This component generates a prompt vector $\mathbf{c}_{G_{\text{P}}}$ for a given product G_{P} based on its search results $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$. This prompt is then injected into the base generative model. We begin by explaining why we prefer encoding the retrieved molecules rather than directly feeding them into the base generative model. Next, we present the neural architecture designed to extract useful clues from these encoded molecules.

Molecule Representation. In estimating $p(G_{\text{R}}|G_{\text{P}})$, a natural solution is to treat G_{P} as a conditioning factor and feed it into a parametric generative model such as the denoiser of a diffusion model D_θ. In the t-th step of its denoising process, we denote the current state by $G^{(t)}_{\text{R}}$ to emphasize the eventual target is the reactant, and $D_\theta(G^{(t)}_{\text{R}}, G_{\text{P}})$ is encouraged to predict G_{R} (Laabid et al. 2024). When conditioned on G_{P}, "alignment" between the reactant and the product is necessary for a permutation-equivariant D_θ to express identity reaction, where the "alignment" has been realized via either Markov bridge (Igashov et al. 2024) or explicit atom mappings (Laabid et al. 2024).

When augmented with $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$, a natural solution is to input all k molecular graphs directly into D_θ, thus preserving as much of the retrieved information as possible. However, unless each $G_i \in \mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}}(G_{\text{P}})$ is "aligned" to the reactant, a permutation-equivariant $D_\theta(G^{(t)}_{\text{R}}, G_{\text{P}}, G_1, \ldots, G_k)$ may not be expressive enough to solve what we define as "union reaction", a simple and necessary capability. Due to the limited space, we defer our detailed analysis to Appendix A. If we follow Laabid et al. (2024) to achieve alignment by the atom mapping matrices, we would need to perform k graph matchings for each G_{P}, which is unaffordable in practice.

Thus, we encode $\mathcal{S}^{(k)}_{\mathcal{D}_{\text{ext}}} = \{G_i\}_{i=1}^k$ into graph-level representations before further processing. Intuitively, we prefer molecule representation learned in a self-supervised manner so that it would not be biased towards certain kind of task. Specifically, for our RARB instance, we adopt Uni-RXN (Qiang et al. 2023), a contrastive learning-based method, and apply its permutation-invariant encoder $\mathcal{E}(\cdot)$ to transform each G_i into an embedding $\mathbf{h}_i$.

Prompt Extraction. Recall that our objective is to extract common structural information from $\{G_i\}_{i=1}^k$ as clues about which parts of G_{P} remain unchanged. To achieve this, we design a prompt extractor that is fed with $\{\mathbf{h}_i\}_{i=1}^k$ and converts them into a single prompt vector $\mathbf{c}_{G_{\text{P}}}$.

Intuitively, mining the common structural information among k molecules requires carefully comparing them pairwise and then aggregating results from each of them. Hence, we parameterize our prompt extractor f_ϕ primarily using a multi-head self-attention block, followed by a sum pooling layer and a linear layer to transform the token embeddings into the final prompt vector.

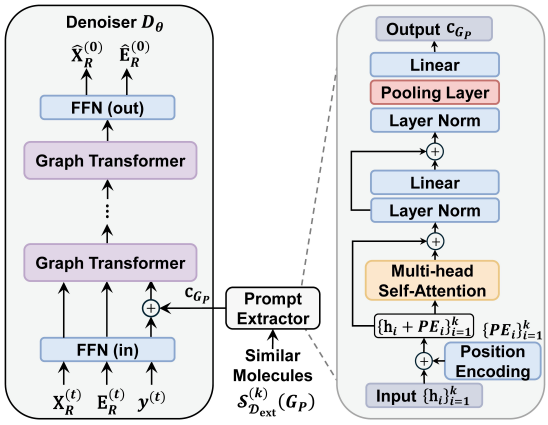

Figure 3: The prompt extractor transforms the embeddings of retrieved molecules into a prompt vector, which is then added to the global features of the graph at current step.

Additionally, we posit that the rankings of these k molecules can reflect their contributions to identifying the unchanged subgraph(s) in the given product, as these rankings are determined by their similarity to the product. Thus, we apply a linear layer to transform the rankings into positional encodings $\{PE_i\}_{i=1}^k$, which are then added to the molecule embeddings to form the initial token embeddings $\{\mathbf{h}_i + PE_i\}_{i=1}^k$. These token embeddings are fed into our prompt extractor to produce a d_{p}-dimensional prompt vector, i.e., $\mathbf{c}_{G_{\text{P}}} = f_\phi(\{\mathbf{h}_i + PE_i\}_{i=1}^k)$. The entire neural architecture is visualized in Fig. 3.

We conduct a preliminary experiment to validate the effectiveness of our prompt extractor parameterization:
Pilot Experiment 3: We compare our design for f_ϕ with an alternative that directly applies an multi-layer perceptron (MLP) to the concatenation $[\mathbf{h}_1, \ldots, \mathbf{h}_k]$. When incorporated with the same base generative model, our design achieves a 65.3% relative improvement in top-1 accuracy over the MLP-based approach, demonstrating its superiority in capturing common substructure information.

More details are provided in Appendix C.3.

3.3 Base Generative Model

We adopt RetroBridge (Igashov et al. 2024), a state-of-the-art method, as the base generative model for RARB. However, our framework is flexible enough to incorporate the retrieval augmentation into any parametric conditional generative model.

Modeling. RetroBridge is built on the Markov bridge framework (Çetin and Danilova 2016), which is particularly well-suited for estimating a conditional distribution, such as $p(G_{\text{R}}|G_{\text{P}})$ in retrosynthesis, from samples drawn from the joint distribution, e.g., $\mathcal{D}_{\text{obs}} = \{(G_{\text{P}_i}, G_{\text{R}_i})\}_{i=1}^{|\mathcal{D}_{\text{obs}}|}$. Unlike traditional diffusion models, RetroBridge replaces the sampling-friendly initial distribution (e.g., Uniform) by the marginal distribution $p(G_{\text{P}})$. Consequently, the denoising process starts with $G_{\text{R}}^{(t)} = G_{\text{P}}$ at $t = T$, where we use $G_{\text{R}}^{(t)}$ to denote the state at timestep t out of a total of T steps. At each step, the ground-truth reactant is predicted by a parametric denoiser based on the current state, i.e., $\hat{G}_{\text{R}}^{(0)} = D_\theta(G_{\text{R}}^{(t)})$. The predicted reactant is then used for sampling the state of the next step $G_{\text{R}}^{(t-1)}$, and this routine continues until $t = 0$.

In RetroBridge, the arguments of the denoiser often include conditioning factors such as G_{P} and/or some (graph-level) global features $\mathbf{y}^{(t)}$ of the current state $G_{\text{R}}^{(t)}$. For RARB, it is natural to include the prompt extracted from G_{P}'s retrieval results as the additional input, i.e., $D_\theta(G_{\text{R}}^{(t)}, G_{\text{P}}, \mathbf{y}^{(t)}, \mathbf{c}_{G_{\text{P}}})$. As shown in Fig. 3, we use a Graph Transformer-based neural architecture (Dwivedi and Bresson 2021; Vignac et al. 2023) as the backbone of our denoiser, maintaining consistency with RetroBridge. Additionally, we add the extracted prompt $\mathbf{c}_{G_{\text{P}}}$ to the global features $\mathbf{y}^{(t)}$ before processing them through the Transformer layers.

Optimization. Generally, diffusion models are optimized by maximizing the Variational Lower Bound (VLB) of the original intractable log-likelihood function. However, DiGress (Vignac et al. 2023) simplifies this by directly minimizing the Cross-Entropy (CE) loss between predicted state at $t = 0$ and the ground-truth target. While RetroBridge prefers VLB and has observed its advantages over CE in retrosynthesis prediction tasks, we consider both options in RARB and compare them in Sec. 4, revealing that each has its own strengths.

As a retrieval-augmented model, RARB faces the risk that the denoiser $D_\theta(G_{\text{R}}^{(t)}, G_{\text{P}}, \mathbf{y}^{(t)}, \mathbf{c}_{G_{\text{P}}})$ might rely too heavily on the provided prompt $\mathbf{c}_{G_{\text{P}}}$. On one hand, we expect this prompt to contain rich information that aids in accurately predicting the correct G_{R}, which should enhance the accuracy of conditional generation. On the other hand, such a helpful conditioning factor might encourage D_θ to use it as a shortcut, potentially leading to an under-utilization of the current state $G_{\text{R}}^{(t)}$.

It is worth noticing that the prompt $\mathbf{c}_{G_{\text{P}}}$ is determined solely by the given product G_{P} and remains fixed throughout the entire denoising process, while the state $G_{\text{R}}^{(t)}$ is stochastic. Over-reliance on the former while neglecting the latter could reduce the diversity of denoising results, resulting in a very limited number of distinct reactants for a given product.

To mitigate the potential decrease in diversity caused by retrieval, we introduce two strategies to increase the randomness ingrained in $\mathbf{c}_{G_{\text{P}}}$: **S1:** to use a dropout rate in the prompt extractor f_ϕ that is higher than that of D_θ. **S2**: to retrieve the top-$(k+e)$ similar molecules through our retrieval component and then randomly select k out of them as the retrieval results $\mathcal{S}_{\mathcal{D}_{\text{ext}}}^{(k)}(G_{\text{P}})$. The extra number of retrieved molecules e is a hyper-parameter, which controls how we balance accuracy and diversity. The second strategy can be consistently applied during both training and inference stages.

4 Experiments

To evaluate the retrieval augmentation introduced in our framework, including both its benefits and potential draw-

backs, we conducted quantitative evaluations to address the following research questions: **RQ1**: Can the retrieval augmentation enhance the performance of the base generative model? **RQ2**: Does it improve out-of-distribution generalization? **RQ3**: Do smaller diffusion models or those with fewer denoising steps still benefit from this augmentation? **RQ4**: Does the retrieval stage negatively impact diversity and efficiency? **RQ5**: Does the quality of retrieved results impact the base generative model? Due to space constraints, the qualitative case study is provided in Appendix D.

4.1 Experimental Setup

Datasets. We conduct our experiments using the USPTO-50k dataset (Schneider, Stiefl, and Landrum 2016), adhering to the standard train/validation/test splits (Dai et al. 2019; Somnath et al. 2021). To assess RARB's ability to handle out-of-distribution data, we construct a more challenging dataset by applying a cluster splitting strategy (Zheng et al. 2019) to USPTO-50k. Specifically, we first use Morgan fingerprints to measure the scaffold similarities between products and employ the Butina algorithm (Butina 1999) to cluster them with a similarity threshold of 0.6. Then, we sort the clusters in descending order by size, and we split these sorted clusters into train/validation/test splits with an 8/1/1 ratio, resulting in our *USPTO-50K-cluster* dataset. For the external dataset, we utilize data from USPTO 2001-2016 applications, comprising $1{,}939{,}254$ raw reactions. We process these raw data using the method proposed by Dai et al. (2019). Importantly, we then remove duplicates, empty reactants, and reactants also present in USPTO-50k, resulting in a final retrieval dataset of $969{,}307$ molecules.

Baselines. We compare RARB with representative methods across three categories: template-based, semi-template, and template-free approaches. Since RARB is a template-free method that incorporates RetroBridge as a component, the comparison will primarily focus on RetroBridge and RARB. Additionally, we evaluate RARB using two commonly employed loss functions: Cross-Entropy and the Variational Lower Bound (VLB).

We select RetroBridge, based on the Graph Transformer architecture, as the base model primarily due to its state-of-the-art performance. We also note that another recent work, DiffAlign (Laabid et al. 2024), which shows comparable performance, is based on a GNN architecture. Our framework is general enough to incorporate various neural architectures, e.g., injecting our encoded retrieval results to a GNN via a virtual node connected to all atoms. Due to the limited space, we will show consistent improvements across multiple base models to further validate our framework in the future.

Metrics. For each input product, we sample 100 reactant sets from RARB and rank them based on their confidence scores, determined by the frequency of their occurrences. We then report *top-k (exact match) accuracy*, which is the proportion of input products for which the model successfully generates the correct reactants within its top-k distinct samples. Next, we use the forward reaction prediction model Molecular Transformer (Schwaller et al. 2019) to predict the products of these top-k samples. We report *round-trip accuracy* (Schwaller et al. 2020) as the percentage of correctly predicted reactants among all predictions. Predicted reactants are considered correct if they either match the ground truth or, when processed by the forward prediction model, lead back to the original input product. We also report *round-trip coverage*, which measures whether there is at least one correct prediction among the top-k samples. Moreover, we define *diversity* as the number of distinct reactants within the 100 samples and report the average diversity value for the test products.

Implementation. We implement RARB based on the open-sourced code of RetroBridge. More details are deferred to Appendix B. Our code is available at this repository: https://github.com/anjie-qiao/RARB.

Model	***k*=1**	**3**	**5**	**10**
Template-Based				
GLN (Dai et al. 2019)	52.5	69.0	75.6	83.7
LocalRetro (Chen and Jung 2021)	**53.4**	**77.5**	**85.9**	**92.4**
Semi-Template				
MEGAN (Sacha et al. 2021)	48.0	70.9	78.1	**85.4**
G2G (Shi et al. 2020)	48.9	67.6	72.5	75.5
RetroXpert (Yan et al. 2020)	50.4	61.1	62.3	63.4
RetroPrime (Wang et al. 2021)	51.4	70.8	74.0	76.1
Retrodiff (Wang et al. 2023)	52.6	**71.2**	**81.0**	83.3
GraphRetro (Somnath et al. 2021)	**53.7**	68.3	72.2	75.5
Template-Free				
SCROP (Zheng et al. 2019)	43.7	60.0	65.2	68.7
Tied Transformer (Kim et al. 2021)	47.1	67.1	73.1	76.3
Aug. Transformer (Tetko et al. 2020)	48.3	-	73.4	77.4
GTA_aug (Seo et al. 2021)	51.1	67.6	74.8	81.6
Graph2SMILES (Tu and Coley 2022)	52.9	66.5	70.0	72.9
Retroformer (Wan et al. 2022)	52.9	68.2	72.5	76.4
DualTF_aug (Sun et al. 2021)	53.6	70.7	74.6	77.0
DiffAlign (Laabid et al. 2024)	54.7	73.3	77.8	81.8
RetroBridge (Igashov et al. 2024)	50.8	74.1	80.6	85.6
RARB (VLB)	56.2	**77.4**	**82.4**	**86.1**
RARB (Cross-Entropy)	**58.3**	75.2	78.8	81.5

Table 1: Top-k (exact match) accuracy on the standard test split of USPTO-50k. The best-performing methods in each group are highlighted in bold.

4.2 Results and Analysis

(RQ1) Improving Base Generative Model. We compare the top-k (exact match) accuracy of RARB with baselines

Model	Coverage			Accuracy		
	k=1	3	5	k=1	3	5
Template-Based						
GLN	82.5	92.0	94.0	82.5	71.0	66.2
LocalRetro	82.1	92.3	94.7	82.1	71.0	66.7
Template-Free						
RetroBridge	84.2	94.3	95.9	84.2	71.7	66.3
RARB	**85.2**	**95.1**	**96.7**	**85.2**	**72.7**	**67.5**

Table 2: Top-k round-trip coverage and accuracy on the standard test split of USPTO-50k.

Model	*k*=1	3	5	10
RetroBridge	42.9	66.3	73.2	77.9
RARB	**54.9**	**71.7**	**76.5**	**78.4**

Table 3: Top-k (exact match) accuracy on the USPTO-50k-cluster dataset's test split.

on the standard test split of USPTO-50k, as shown in Table 1. RARB using VLB outperforms other methods across all top-k accuracy, while the RARB using Cross-Entropy achieves higher top-1 accuracy, even surpassing template-based methods. Given the significantly faster training and convergence, we choose Cross-Entropy as our loss function for the remaining experiments.

Next, we compare round-trip coverage and accuracy of RARB and its base generative model, as well as template-based methods, as shown in Table 2. RARB not only outperforms its base generative model but also surpasses state-of-the-art template-based methods in all top-*k* round-trip metrics. These results indicate that the retrieval augmentation introduced by our framework enhances the performance of the base generative model.

(RQ2) Out-of-distribution Generalization. Given that the real-world chemical space is vastly larger than the space covered by the training set, the model's ability to handle out-of-distribution (OOD) data becomes particularly critical. Thus, we compare RARB with RetroBridge on the USPTO-50k-cluster dataset, which was divided based on scaffold similarity clustering to effectively simulate OOD scenarios. Retrosynthesis prediction on such a test split is more challenging, as the reactions in the validation and test sets exhibit lower similarity to those in the training set.

As shown in Table 3, RARB's performance on the USPTO-50k-cluster only slightly decreases compared to that on the standard splits, while its improvement relative to its base generative model, RetroBridge, becomes more pronounced. Notably, in terms of top-1 accuracy, the relative improvement increases to 28%. This result demonstrates that our framework enhances its base generative model's ability to handle OOD molecules.

(RQ3) Performance with Reduced Model Complexity. The retrieved information functions as a non-parametric memory, providing additional context to guide the generation process and helping to bridge gaps in the model's learned knowledge. This approach has the potential to enhance the performance of generative models in resource-constrained scenarios. To validate this, we train RARB in two conditions: **(i)** using only about 60% of the base model's parameters, and **(ii)** reducing the base generative model's denoising steps from 500 to 200. The results are shown in Table 4. Even with fewer parameters or denoising steps, RARB outperforms the base generative model across all top-*k* accuracy metrics. Notably, reducing the denoising steps by 60% has only a minimal impact on RARB. These results highlight RARB's advantage in scenarios with limited resources or where efficient inference is necessary. These advantages are particularly valuable in retrosynthesis, where high-throughput prediction is crucial for efficiently exploring large chemical spaces and accelerating discovery.

Model	#param	#step	*k*=1	3	5	10
RetroBridge	4.8M	500	50.8	71.1	76.0	80.3
RARB	**2.8M**	500	54.9	73.7	78.3	**82.2**
	4.8M	**200**	**57.2**	**74.9**	**79.0**	82.0

Table 4: Top-k (exact match) accuracy with reduced model complexity.

(RQ4) Diversity and Efficiency. The evaluations above highlight RARB's advantages in enhancing the base generative model. Here, we investigate the potential impacts and additional overhead that retrieval-augmented strategy may introduce to the base generative model.

(1) *Diversity.* Diversity is crucial for generative models in real-world applications, especially in molecule generation, which involves exploring the broad chemical space. As discussed in Sec. 3.3, we are concerned that the denoiser's over-reliance on retrieval results may reduce the diversity of its generated samples.

As shown in Table 5, RARB, without any specific strategy to improve diversity, generates molecules that are less diverse than those generated by its base generative model, confirming our concern. Increasing the dropout rate for prompt extractor to 0.5 (i.e., our first strategy (S1)) not only enhances RARB's diversity but also improves its accuracy, supporting our rationale of reducing the denoiser's reliance on shortcuts. Additionally, we further add our second strategy (S2) to RARB, namely, randomly selecting 3 out of the top-5 ranked molecules as the retrieval results. These two strategies, we proposed in Sec. 3.3, together lead to a 60.7% increase in diversity compared to RARB without these strategies, while also delivering better top-k accuracy.

(2) Efficiency. While retrieving additional information and extracting prompt for the base generative model does add some time overhead, our evaluation shows that this overhead is negligible compared to the denoising process of RetroBridge. Specifically, the retrieval of the top-10 similar molecules from the USPTO-applications dataset and their conversion into embeddings takes approximately 2 seconds per sample, resulting in only a 1.9% increase in overall inference time. This slight increase is expected, given that the time-intensive sampling in diffusion models far outweighs the retrieval process.

Model	Diversity	Top-*k* accuracy				Strategy
		$k=1$	3	5	10	
RetroBridge	**12.9**	50.8	71.1	76.0	80.3	-
RARB	7.47	**58.3**	71.5	74.1	75.7	w/o
	10.1	**58.3**	75.2	78.8	81.5	w/ S1
	12.0	56.0	**76.4**	**81.6**	**84.9**	w/ S1&S2

Table 5: Comparisons of diversity on USPTO-50k, where S1 means dropout=0.5 for our prompt extractor, and S2 means sampling 3 out of top-5 molecules as retrieval results.

Average similarity	k=1	3	5	10
75%∼100% (most similar)	4.8	2.5	1.5	-0.1
50%∼75%	12.3	6.1	4.5	3.0
25%∼50%	12.2	8.3	5.9	4.2
0%∼25% (least similar)	0.8	-0.6	-0.9	-2.4

Table 6: Relative improvements of RARB over the base generative model in Top-k (exact match) accuracy across different similarity groups of the test set.

(RQ5) Influence of Retrieval Quality. The retrieval-augmented strategy may raise a potential concern: how the relevance between the query and the retrieved information impacts the performance of the base generative model. Here, we assess the effect of varying similarity levels between the product and retrieved molecules, focusing particularly for those with low similarity or complex molecules.

To do this, We rank the standard test set by the average similarity between the product and the retrieved top-3 molecules, in ascending order. We then divide the set into four equal groups. Finally, we compute the relative improvements of RARB over RetroBridge for each groups, as shown in Table 6.

The results suggest that the benefits of retrieved molecules for the base model generally increase with higher similarity. In the least similar group, accuracy slightly declined, likely due to irrelevant noise introduced by low-similarity retrievals. Moreover, in the most similar group, the improvements become less significant, as highly similar molecules may limit the diversity of functional groups, potentially overlooking those crucial for the reactants.

In Sec. 3.3, we propose the S1 and S2 strategies to promote diversity in generative model sampling. Here, we incorporate these strategies into RARB and evaluate the relative improvements across these groups. As shown in Table 7, the results demonstrate that, with these strategies, even the least similar group shows positive benefits. This suggests that our strategies further enhance the model's robustness to low-similarity retrievals.

In reality, the reaction centers of product molecules may involve multiple bonds or atoms, referred to multiple reaction centers. These products often have complex structures, and the retrieved molecules may vary significantly. In the following, we focus on the performance of RARB on such complex products.

We adopt the scheme from (Wan et al. 2022) to identify

Average similarity	k=1	3	5	10
75%∼100% (most similar)	2.8	4.4	5.1	3.4
50%∼75%	9.0	6.3	6.3	5.0
25%∼50%	8.4	8.0	7.5	6.4
0%∼25% (least similar)	0.6	2.5	3.3	3.4

Table 7: Relative improvements of RARB with S1&S2 strategies over the base generative model in Top-k (exact match) accuracy across all groups of the test set.

Products	k=1	3	5	10
RC=1	5.5	5.5	5.8	4.8
RC>1	0.4	1.3	1.2	1.2
Full Test Set	5.2	5.3	5.6	4.6

Table 8: Relative improvements of RARB over the base generative model in Top-k (exact match) accuracy across different groups of the test set.

reaction centers (RC) and classify products with more than one RC as complex. In the standard test set, 252 out of 5007 products are categorized as complex. We then calculate the relative improvement of RARB over the base model for both groups, as shown in Table 8. The results show that RARB's improvements on complex products are more modest compared to the full test set, highlighting the need for future work to develop retrieval strategies tailored to the unique characteristics of complex molecules.

5 Conclusion

In this work, we propose a RAG framework for template-free, single-step retrosynthesis prediction, driven by our key insight that similar molecules can provide valuable clues about the invariant substructures during chemical reactions. We instantiate our framework as RARB, which significantly enhances the performance of its base generative model, RetroBridge, across all relevant metrics. Particularly, RARB's superiority in handling out-of-distribution molecules addresses a critical need in real-world applications.

We believe that the success of RARB paves the way for further applications of RAG within drug discovery and a broader range of molecule-centric scientific tasks. Meanwhile, as the improvements introduced by RAG tend to heavily depend on the quality of retrieval, a novel fragment/molecule retrieval scheme would be an important research topic. Besides, the risk of data leakage should be carefully considered, which might require more fair evaluation protocol for RAG-based generative models.

Acknowledgments

This research was supported in part by National Natural Science Foundation of China (No. 92470128, No. U2241212, No. 61932001). Zhen Wang was supported by National Natural Science Foundation of China under Grant 62302537, Guangzhou Basic and Applied Basic Research Foundation under Grant 2024A04J4449, and Fundamental Research Funds for the Central Universities, Sun Yat-sen University, under Grant 24qnpy139. This work is conducted in part on RTAI cluster, which is supported by School of Computer Science and Engineering and Institute of Artificial Intelligence, Sun Yat-sen University. We would like to thank the anonymous reviewers for their insightful comments and suggestions, which significantly improved our paper's quality.

References

Austin, J.; Johnson, D. D.; Ho, J.; Tarlow, D.; and van den Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 17981–17993.

Bajusz, D.; Rácz, A.; and Héberger, K. 2015. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7: 1–13.

Blakemore, D. C.; Castro, L.; Churcher, I.; Rees, D. C.; Thomas, A. W.; Wilson, D. M.; and Wood, A. 2018. Organic synthesis provides opportunities to transform drug discovery. *Nature Chemistry*, 10(4): 383–394.

Blattmann, A.; Rombach, R.; Oktay, K.; Müller, J.; and Ommer, B. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 15309–15324.

Butina, D. 1999. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *Journal of Chemical Information and Computer Sciences*, 39: 747–750.

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. 2021. Extracting training data from large language models. In *USENIX Security Symposium*, 2633–2650.

Çetin, U.; and Danilova, A. 2016. Markov bridges: SDE representation. *Stochastic Processes and their Applications*, 126(3): 651–679.

Chen, S.; and Jung, Y. 2021. Deep retrosynthetic reaction prediction using local reactivity and global attention. *Journal of the American Chemical Society Au*, 1(10): 1612–1620.

Dai, H.; Li, C.; Coley, C.; Dai, B.; and Song, L. 2019. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 8780–8794.

Dwivedi, V. P.; and Bresson, X. 2021. A Generalization of Transformer Networks to Graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications*.

Fang, L.; Li, J.; Zhao, M.; Tan, L.; and Lou, J.-G. 2023. Single-step retrosynthesis prediction by leveraging commonly preserved substructures. *Nature Communications*, 14(1): 2446.

Han, Y.; Xu, X.; Hsieh, C.-Y.; Ding, K.; Xu, H.; Xu, R.; Hou, T.; Zhang, Q.; and Chen, H. 2024. Retrosynthesis prediction with an iterative string editing model. *Nature Communications*, 15(1): 6404.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 6840–6851.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 8633–8646.

Hoogeboom, E.; Satorras, V. G.; Vignac, C.; and Welling, M. 2022. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning (ICML)*, 8867–8887.

Igashov, I.; Schneuing, A.; Segler, M.; Bronstein, M.; and Correia, B. 2024. RetroBridge: Modeling Retrosynthesis with Markov Bridges. In *International Conference on Learning Representations (ICLR)*.

Kang, M.; Gürel, N. M.; Yu, N.; Song, D.; and Li, B. 2024. C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models. In *International Conference on Machine Learning (ICML)*.

Kim, E.; Lee, D.; Kwon, Y.; Park, M. S.; and Choi, Y.-S. 2021. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *Journal of Chemical Information and Modeling*, 61(1): 123–133.

Kong, Z.; Ping, W.; Huang, J.; Zhao, K.; and Catanzaro, B. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. In *International Conference on Learning Representations (ICLR)*.

Laabid, N.; Rissanen, S.; Heinonen, M.; Solin, A.; and Garg, V. 2024. Alignment is Key for Applying Diffusion Models to Retrosynthesis.

Li, X.; Thickstun, J.; Gulrajani, I.; Liang, P. S.; and Hashimoto, T. B. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 4328–4343.

Liu, S.; Tu, Z.; Xu, M.; Zhang, Z.; Lin, L.; Ying, R.; Tang, J.; Zhao, P.; and Wu, D. 2023. FusionRetro: molecule representation fusion via in-context learning for retrosynthetic planning. In *International Conference on Machine Learning (ICML)*, 22028–22041.

Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Association for Computational Linguistics (ACL)*, 9802–9822.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 4172–4182.

Qiang, B.; Zhou, Y.; Ding, Y.; Liu, N.; Song, S.; Zhang, L.; Huang, B.; and Liu, Z. 2023. Bridging the gap between chemical reaction pretraining and conditional molecule generation with a unified model. *Nature Machine Intelligence*, 5(12): 1476–1485.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Sacha, M.; Błaz, M.; Byrski, P.; Dabrowski-Tumanski, P.; Chrominski, M.; Loska, R.; Włodarczyk-Pruszynski, P.; and Jastrzebski, S. 2021. Molecule edit graph attention network: modeling chemical reactions as sequences of graph

edits. *Journal of Chemical Information and Modeling*, 61(7): 3273–3284.

Schneider, N.; Stiefl, N.; and Landrum, G. A. 2016. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling*, 56(12): 2336–2346.

Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; and Lee, A. A. 2019. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9): 1572–1583.

Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; and Laino, T. 2020. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science*, 11(12): 3316–3325.

Segler, M. H.; Preuss, M.; and Waller, M. P. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698): 604–610.

Seo, S.-W.; Song, Y. Y.; Yang, J. Y.; Bae, S.; Lee, H.; Shin, J.; Hwang, S. J.; and Yang, E. 2021. GTA: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, 531–539.

Shi, C.; Xu, M.; Guo, H.; Zhang, M.; and Tang, J. 2020. A graph to graphs framework for retrosynthesis prediction. In *International Conference on Machine Learning (ICML)*, 8818–8827.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2256–2265.

Somnath, V. R.; Bunne, C.; Coley, C.; Krause, A.; and Barzilay, R. 2021. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 9405–9415.

Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H.; and Glorius, F. 2020. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews*, 49(17): 6154–6168.

Sun, R.; Dai, H.; Li, L.; Kearnes, S.; and Dai, B. 2021. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 10186–10194.

Tetko, I. V.; Karpov, P.; Van Deursen, R.; and Godin, G. 2020. State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Communications*, 11(1): 5575.

Tu, Z.; and Coley, C. W. 2022. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of Chemical Information and Modeling*, 62(15): 3503–3513.

Vignac, C.; Krawczuk, I.; Siraudin, A.; Wang, B.; Cevher, V.; and Frossard, P. 2023. Digress: Discrete denoising diffusion for graph generation. In *International Conference on Learning Representations (ICLR)*.

Wan, Y.; Hsieh, C.-Y.; Liao, B.; and Zhang, S. 2022. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *International Conference on Machine Learning (ICML)*, 22475–22490.

Wang, X.; Li, Y.; Qiu, J.; Chen, G.; Liu, H.; Liao, B.; Hsieh, C.-Y.; and Yao, X. 2021. RetroPrime: A Diverse, plausible and Transformer-based method for Single-Step retrosynthesis predictions. *Chemical Engineering Journal*, 420: 129845.

Wang, Y.; Song, Y.; Xu, M.; Wang, R.; Zhou, H.; and Ma, W. 2023. RetroDiff: Retrosynthesis as Multi-stage Distribution Interpolation.

Weininger, D. 1988. 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.

Xu, R.; Guo, M.; Wang, J.; Li, X.; Zhou, B.; and Loy, C. C. 2021. Texture memory-augmented deep patch-based image inpainting. *IEEE Transactions on Image Processing*, 30: 9112–9124.

Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; and Huang, J. 2020. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 11248–11258.

Zeng, K.; Yang, B.; Zhao, X.; Zhang, Y.; Nie, F.; Yang, X.; Jin, Y.; and Xu, Y. 2024. Ualign: pushing the limit of template-free retrosynthesis prediction with unsupervised SMILES alignment. *Journal of Cheminformatics*, 16(1): 80.

Zhang, X.; Mo, Y.; Wang, W.; and Yang, Y. 2024. Retrosynthesis prediction enhanced by in-silico reaction data augmentation.

Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; and Cui, B. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey.

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

Zhong, W.; Yang, Z.; and Chen, C. Y.-C. 2023. Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nature Communications*, 14(1): 3009.

Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Hou, T.; and Song, M. 2023. Recent advances in artificial intelligence for retrosynthesis.

Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; and Song, M. 2022. Root-aligned SMILES: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31): 9023–9034.

Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; and Ke, G. 2023. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. In *International Conference on Learning Representations (ICLR)*.