
RetroDiff: Retrosynthesis as Multi-stage Distribution Interpolation

Yiming Wang^{1,*}, Yuxuan Song², Yiqun Wang³, Minkai Xu⁴, Rui Wang¹, Hao Zhou^{2,†}, Wei-Ying Ma²

¹Department of Computer Science and Engineering, Shanghai Jiao Tong University

²Institute for AI Industry Research (AIR), Tsinghua University

³ByteDance Research ⁴Department of Computer Science, Stanford University

*Work done during Yiming’s internship at AIR, Tsinghua University. †Corresponding Author

Abstract

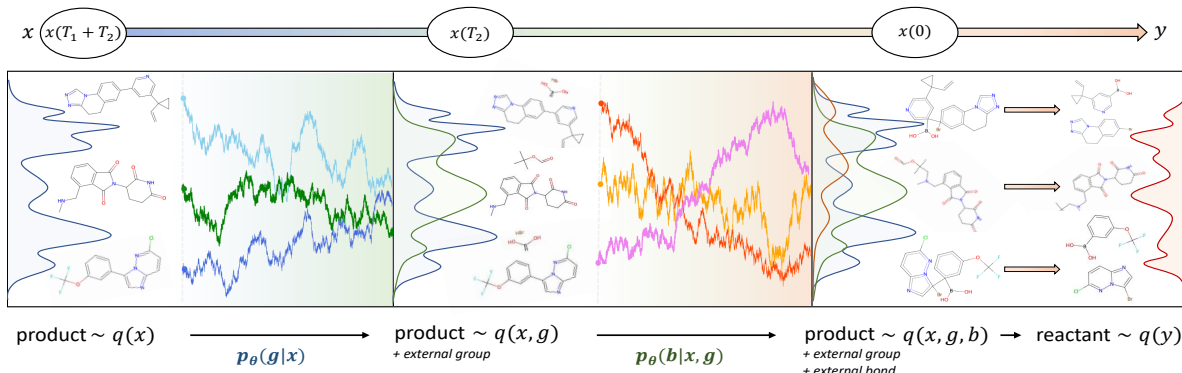
Retrosynthesis poses a key challenge in biopharmaceuticals, aiding chemists in finding appropriate reactant molecules for given product molecules. With reactants and products represented as 2D graphs, retrosynthesis constitutes a conditional graph-to-graph (G2G) generative task. Inspired by advancements in discrete diffusion models for graph generation, we aim to design a diffusion-based method to address this problem. However, integrating a diffusion-based G2G framework while retaining essential chemical reaction template information presents a notable challenge. Our key innovation involves a multi-stage diffusion process. We decompose the retrosynthesis procedure to first sample external groups from the dummy distribution given products, then generate external bonds to connect products and generated groups. Interestingly, this generation process mirrors the reverse of the widely adapted semi-template retrosynthesis workflow, *i.e.* from reaction center identification to synthon completion. Based on these designs, we introduce Retrosynthesis Diffusion (RetroDiff), a novel diffusion-based method for the retrosynthesis task. Experimental results demonstrate that RetroDiff surpasses all semi-template methods in accuracy, and outperforms template-based and template-free methods in large-scale scenarios and molecular validity, respectively. Code: <https://github.com/Alsace08/RetroDiff>.

1 Introduction

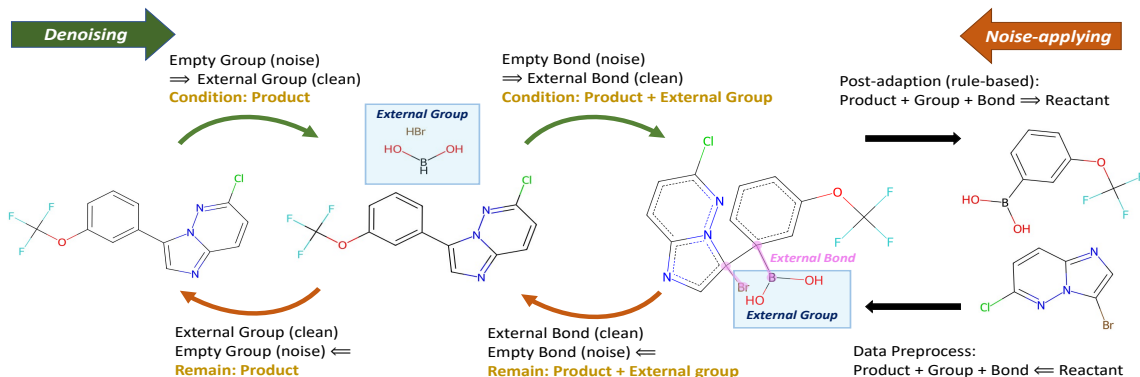
Retrosynthesis (Corey, 1991) is important in organic synthesis, which helps chemists find legitimate reactant molecules given product molecule, thus providing efficient and stable drug discovery and compound preparation methods for the biopharmaceutical field. Since the first computer-aided method was proposed (Corey & Wipke, 1969), huge efforts have been devoted to exploring analytical computational methods for retrosynthesis, and research for data-driven methods has reached its peak in recent years.

Retrosynthesis methods can be broadly categorized into three groups. *Template-based* methods retrieve the best match reaction template for a target molecule from large-scale chemical databases (Schneider et al., 2016; Chen & Jung, 2021). Though with appealing performance, the scalability of template-based methods is indeed limited by the template database size (Segler & Waller, 2017a; Segler et al., 2018); *Template-free* methods generate the reactants given corresponding products directly without any chemical prior (Zheng et al., 2019; Seo et al., 2021; Tu & Coley, 2022), but limited chemical reaction diversity and interpretability hinder the potential of them in practical applications (Chen et al., 2019; He et al., 2018; Jiang & de Rijke, 2018; Roberts et al., 2020).

Fortunately, *semi-template* methods could be another alternative for building retrosynthesis models. Combining the strengths of both template-based and -free methods, semi-template methods introduce the chemical prior into models by employing a two-stage process including “reaction center prediction” and “synthon completion”. This makes semi-template methods more scalable than template-based ones and more interpretable than template-free ones, which has drawn increasing interest of late (Yan et al., 2020; Shi et al., 2020; Wang et al., 2021). In this paper, we aim to develop a more effective semi-template method.



(a) RetroDiff Pipeline: Macro Denoising Process



(b) RetroDiff Example: Micro Noise-applying and Denoising Process

Figure 1: The pipeline and example of our RetroDiff model.

The non-autoregressive generative model diffusion is particularly well-suited for capturing the complex structure of graph data, along with its robust capability for probabilistic modeling. However, the intractable reactant and product distributions impede a naive adoption of diffusion models to smoothly interpolate between these chemical spaces. Moreover, current reaction templates overly constrain the intrinsic data structure and necessitate artificial modifications to the molecular structure of groups and bonds, making it difficult to provide the explicit product prior for the diffusion modeling. To address this issue, we redefine the reaction template by separating the external group generation from the external bond generation. This revised approach aligns with the concept of retrosynthesis, wherein the task is to transform distributions with minimal constraints: *given a product molecule, we generate a dummy distribution that transitions to distributions of external groups and bonds, then we splice these to form the reactant distribution*. With such a template setup, we cleverly assign the intractable product distribution to learning conditions rather than goals via chemical prior.

Building on this template, we introduce **RetroDiff** — a **Retrosynthesis Diffusion** model that works in discrete conditions, as illustrated in Figure 1. The model first generates molecular structures through a two-stage

denoising process: Initially, it begins with a prior distribution, proceeding first to create the external groups, which are parts that attach to the product molecule (Sec.2.1.1); Once these groups are formed, the model then constructs the bonds that connect these external groups to the product (Sec.2.1.2). Finally, We manually remove some product bonds based on the reaction sites identified by the generated external bonds, thereby ensuring the resulting reactant is chemically valid (Sec.2.1.3). RetroDiff innovatively flips the script on the conventional semi-template methods: In our method, the high-uncertainty variables (groups) are first generated, this significantly minimizes the error buildup of generating low entropy variables (bonds).

We conduct extensive experiments (Sec.3.2) on the USPTO-50k (Schneider et al., 2016) and USPTO-full (Lowe, 2017) dataset, and empirical results show that our model achieves **state-of-the-art top-k performance** compared with other competitive *semi-template* methods. When compared with *template-free* methods, our method is competitive but exhibits a **higher validity** advantage, which ensures our stronger availability and security in real scenarios. When compared with *template-based* methods, our method has a slight performance disadvantage on USPTO-50k. However, when the application scene is further scaled to include a larger chemical space, our

method’s performance far exceeds other template-based methods, which is verified on the large-scale USPTO-full, indicating that we achieve **greater scalability**.

2 RetroDiff: Retrosynthesis Diffusion

We begin by defining the task of retrosynthesis prediction. Consider a chemical reaction expressed as $\{\mathbf{G}_R^i\}_{i=1}^{|R|} \rightarrow \{\mathbf{G}_P^i\}_{i=1}^{|P|}$, where \mathbf{G}_R represents the set of reactant molecular graphs, \mathbf{G}_P represents the set of product molecular graphs, and $|R|$ and $|P|$ indicate the respective counts of reactants and products in a given reaction. Typically, we assume $|P| = 1$, which aligns with the conventions of benchmark datasets. The key problem in the retrosynthesis task is to invert the chemical reaction; namely deduce the reactant set $\{\mathbf{G}_R^i\}_{i=1}^{|R|}$ when presented with a sole product $\{\mathbf{G}_P\}$. In general, the assorted connected sub-graphs comprising the reactants can be amalgamated into a single disjoint graph $\{\mathbf{G}_R\}$. Thus, the retrosynthesis prediction simplifies to the transformation $\{\mathbf{G}_P\} \rightarrow \{\mathbf{G}_R\}$.

Existing semi-template methods typically first identify the reaction center in the given product and then complete the synthons at the fractured site. However, such a template setup is infeasible for designing the appropriate generative diffusion process. To address this, we redefine the task template with the following preliminary notations: $\mathbf{x} \sim P_{\mathcal{X}}$ denotes the variable of product graphs and the corresponding distribution, $\mathbf{y} \sim P_{\mathcal{Y}}$ as the reactant variable, $\mathbf{g} \sim P_{\mathcal{G}}$ as the external group, and $\mathbf{b} \sim P_{\mathcal{B}}$ as the external bond. We elaborate on the redefined template in the following stages:

- **Stage 1: External Group Generation.** The process commences with the generation of external group \mathbf{g} that will attach to the product \mathbf{x} . Namely sampling from such distribution $P_{\mathcal{G}}(\mathbf{g}|\mathbf{x};\theta)$ which is parameterized by the neural network θ .
- **Stage 2: External Bond Generation.** Next, the process involves the generation of the external bond \mathbf{b} , which will link the product \mathbf{x} with the newly formed external group \mathbf{g} . Here, we focus on modeling the distribution $P_{\mathcal{B}}(\mathbf{b}|\mathbf{g},\mathbf{x};\theta)$.
- **Stage 3: Post-Adaptation (Rule-Based).** The concluding phase involves a manual adjustment, breaking the reaction center in the product in line with valence rules to yield the final reactant \mathbf{y} . This transformation is depicted as $P_{\mathcal{Y}}(\mathbf{y}|\mathbf{b},\mathbf{g},\mathbf{x})$ which is a predetermined rule-based mapping.

Building on this framework, we introduce RetroDiff, which integrates the above stages into a unified diffusion model. This serial procedure essentially implies an autoregressive decomposition of the probabilistic model, aimed at approximating the conditional distribution:

$$\begin{aligned} P_{\text{model}}(\mathbf{y}|\mathbf{x};\theta) \\ = \int P_{\mathcal{G}}(\mathbf{g}|\mathbf{x};\theta)P_{\mathcal{B}}(\mathbf{b}|\mathbf{g},\mathbf{x};\theta)P_{\mathcal{Y}}(\mathbf{y}|\mathbf{b},\mathbf{g},\mathbf{x}) \, \text{d}\mathbf{b} \, \text{d}\mathbf{g}, \end{aligned} \quad (1)$$

which essentially represents the transformation between distributions of product and reactant.

2.1 RetroDiff Pipeline

In this section, we introduce the whole pipeline of the proposed RetroDiff which includes the detailed implementations of $P_{\mathcal{G}}(\mathbf{g}|\mathbf{x};\theta)$, $P_{\mathcal{B}}(\mathbf{b}|\mathbf{g},\mathbf{x};\theta)$ and $P_{\mathcal{Y}}(\mathbf{y}|\mathbf{b},\mathbf{g},\mathbf{x})$, as presented in the Eq.1.

We utilize the diffusion process to model all the conditional distributions. For completeness, we elaborate on the details for parameterizing the conditional distribution with a diffusion process. We take $P_{\mathcal{G}}(\mathbf{g}|\mathbf{x};\theta)$ as an example. Under the context of diffusion models, the dimensions of the input and output variables should be aligned. Hence we append a dummy noisy variable \mathbf{v}_1 which makes the input as $(\mathbf{v}_1, \mathbf{x})$; Correspondingly, the output is as (\mathbf{g}, \mathbf{x}) . Note that here we have $\dim(\mathbf{v}_1) = \dim(\mathbf{g})$. Similarly, for $P_{\mathcal{Y}}(\mathbf{y}|\mathbf{b},\mathbf{g},\mathbf{x})$, the input is $(\mathbf{v}_2, \mathbf{g}, \mathbf{x})$ while the output is as $(\mathbf{b}, \mathbf{g}, \mathbf{x})$. For the training objective, we only calculate the objective on the variables concerned, \mathbf{g} in $P_{\mathcal{G}}(\mathbf{g}|\mathbf{x};\theta)$ and \mathbf{b} in $P_{\mathcal{B}}(\mathbf{b}|\mathbf{g},\mathbf{x};\theta)$. Strictly, our model implies a transformation in the joint space as:

$$\mathcal{X} \times \mathcal{V}_1 \times \mathcal{V}_2 \rightarrow \mathcal{X} \times \mathcal{G} \times \mathcal{V}_2 \rightarrow \mathcal{X} \times \mathcal{G} \times \mathcal{B} \rightarrow \mathcal{Y}, \quad (2)$$

Details of the generation pipeline can be found in Figure 2 (Denoising Process). To simplify the representation, we denote the condition at each stage as \mathbf{c} .

2.1.1 External Group Generation

The goal of this stage is to interpolate the distribution $P_{\mathcal{V}_1}$ to $P_{\mathcal{G}}$ conditioned on \mathbf{c} . In this stage, condition \mathbf{c} is the product $\mathbf{x} \sim P_{\mathcal{X}}$. This is a **graph-to-graph** generative process, we define $\mathbf{v} \sim P_{\mathcal{V}_1}$ as a dummy noisy graph and $\mathbf{g} \sim P_{\mathcal{G}}$ as a true external group.

Noise-applying. In the noise-applying process, we interpolate the distribution $P_{\mathcal{G}}$ to $P_{\mathcal{V}_1}$. With a slight abuse of notation, we splice external group \mathbf{g} and product \mathbf{x} into one unconnected graph $\mathbf{G} = (\mathbf{X}, \mathbf{E})$ with n atoms and m bonds, each atom and bond have a and b categories, respectively, so they can be represented by one-hot attributes that $\mathbf{X} \in \mathbb{R}^{n \times a}$ and $\mathbf{E} \in \mathbb{R}^{n \times n \times b}$. For graph \mathbf{G} , each atom and bond are diffused independently (Vignac et al., 2022), which means the state transition each time acts on the single atom $x_i \in \mathbf{X}$ and bond $e_i \in \mathbf{E}$.

We follow (Austin et al., 2021) to define the Markov matrix \mathbf{Q}_t to conduct probability transitions of states

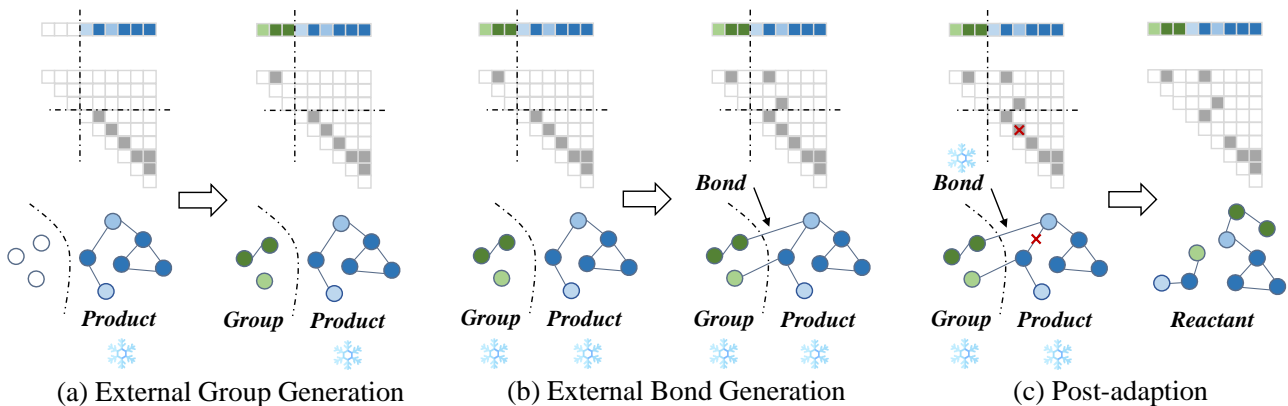


Figure 2: The generation overview of the distribution transformation upon our template. The top row indicates changes in the atom types in the graph, the middle row indicates changes in the adjacency matrix of the graph, and the bottom row indicates overall changes in the graph structure. Specifically, the hollow circle denotes a dummy atom category we set for this task, and the colored circles denote real atom categories. The Line between circles means there exists one bond between the two atoms.

at each time t in the discrete space. For graph \mathbf{G} , we apply noise to atoms via $[\mathbf{Q}_t^X]_{ss'} = q(x_t = s' | x_{t-1} = s)$ and bonds via $[\mathbf{Q}_t^E]_{ss'} = q(e_t = s' | e_{t-1} = s)$, where s and s' represent the atom/bond state at time $t-1$ and t , respectively. Due to the graph independence, the noise-applying process for graph \mathbf{G} can be defined as:

$$\begin{aligned} q(\mathbf{G}_t | \mathbf{G}_{t-1}) &= (\mathbf{X}_{t-1} \mathbf{Q}_t^X, \mathbf{E}_{t-1} \mathbf{Q}_t^E) \\ \implies q(\mathbf{G}_t | \mathbf{G}_0) &= (\mathbf{X}_0 \bar{\mathbf{Q}}_t^X, \mathbf{E}_0 \bar{\mathbf{Q}}_t^E), \end{aligned} \quad (3)$$

where \mathbf{G}_0 is the graph of ground truth, $\bar{\mathbf{Q}}_t^X = \prod_{i=1}^t \mathbf{Q}_i^X$ and $\bar{\mathbf{Q}}_t^E = \prod_{i=1}^t \mathbf{Q}_i^E$. Finally, we sample the probability distribution $q(\mathbf{G}_t | \mathbf{G}_0)$ to obtain the noisy graph \mathbf{G}_t .

Denoising. In the denoising process, given a noisy graph \mathbf{G}_t and condition \mathbf{c} , we need to iterate the denoising process $p_\theta(\mathbf{G}_{t-1} | \mathbf{G}_t, \mathbf{c})$ by a trainable network p_θ at each time t . We model the distribution as the product over nodes and edges and marginalize each item over the network predictions:

$$p_\theta(\mathbf{G}_{t-1} | \mathbf{G}_t, \mathbf{c}) = \prod_{x \in \mathbf{X}_{t-1}} p_\theta(x | \mathbf{G}_t, \mathbf{c}) \prod_{e \in \mathbf{E}_{t-1}} p_\theta(e | \mathbf{G}_t, \mathbf{c}), \quad (4)$$

where

$$\begin{aligned} p_\theta(x | \mathbf{G}_t, \mathbf{c}) &= \sum_{x_0 \in \mathbf{X}_0} q(x | x_t, x_0, \mathbf{c}) p_\theta(x_0 | \mathbf{G}_t, \mathbf{c}), \\ p_\theta(e | \mathbf{G}_t, \mathbf{c}) &= \sum_{e_0 \in \mathbf{E}_0} q(e | e_t, e_0, \mathbf{c}) p_\theta(e_0 | \mathbf{G}_t, \mathbf{c}). \end{aligned} \quad (5)$$

Next, we derive $q(\mathbf{G}_{t-1} | \mathbf{G}_t, \mathbf{G}_0, \mathbf{c})$ with the Bayes theorem and transform it into forms of node and edge to complete the calculations in Eq.4 (Vignac et al.,

2022). For node X , we have:

$$\begin{aligned} q(\mathbf{X}_{t-1} | \mathbf{X}_t, \mathbf{X}_0, \mathbf{c}) &= \frac{q(\mathbf{X}_t | \mathbf{X}_{t-1}, \mathbf{X}_0, \mathbf{c}) q(\mathbf{X}_{t-1} | \mathbf{X}_0, \mathbf{c})}{q(\mathbf{X}_t | \mathbf{X}_0, \mathbf{c})} \\ &= \frac{\mathbf{X}_t [\mathbf{Q}_t^X]^\top \odot \mathbf{X}_0 \bar{\mathbf{Q}}_{t-1}^X}{\mathbf{X}_0 \bar{\mathbf{Q}}_t^X [\mathbf{X}_t]^\top} \propto \mathbf{X}_t [\mathbf{Q}_t^X]^\top \odot \mathbf{X}_0 \bar{\mathbf{Q}}_{t-1}^X. \end{aligned} \quad (6)$$

Similarly, $q(\mathbf{E}_{t-1} | \mathbf{E}_t, \mathbf{X}_0, \mathbf{c}) \propto \mathbf{E}_t [\mathbf{Q}_t^E]^\top \odot \mathbf{E}_0 \bar{\mathbf{Q}}_{t-1}^E$. Based on this derivation, we only need to create a network $p_\theta(\mathbf{G}_0 | \mathbf{G}_t, \mathbf{c})$ to predict clean graph \mathbf{G}_0 given noisy data \mathbf{G}_t and condition \mathbf{c} .

2.1.2 External Bond Generation

In this stage, we aim to interpolate the distribution $P_{\mathcal{V}_2}$ to $P_{\mathcal{B}}$ conditioned on \mathbf{c} , where condition \mathbf{c} is $P_{\mathcal{X}} \times P_{\mathcal{G}}$. This is a **bond-to-bond** generative process, we define $\mathbf{v} \sim P_{\mathcal{V}_2}$ as the dummy noisy bond and $\mathbf{b} \sim P_{\mathcal{B}}$ connecting \mathbf{g} and \mathbf{x} , and splice \mathbf{g} , \mathbf{x} , and \mathbf{b} as a connected graph. We have obtained a trained network p_θ in the last stage, so we freeze \mathbf{g} and \mathbf{x} in the graph and continue to train p_θ . The principles of the noise-applying and denoising processes in this stage are the same as in Section 2.1.1, with the only difference being the spaces at both ends of the interpolation.

2.1.3 Post-adaption

Now we get three parts: product (inherent), external group, and external bond. We need to combine these three into final reactants that are chemically legal.

In the traditional template definition of semi-template methods (Yan et al., 2020; Shi et al., 2020; Wang et al., 2021), bonds in the product are first broken to create reaction sites ("reaction center prediction", usually one bond), and then new leaving groups are generated on them ("synthon completion"). Back to our steps, we

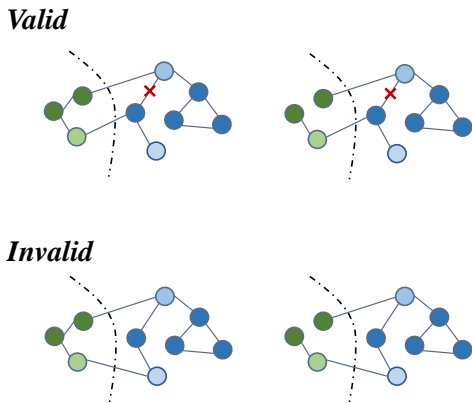


Figure 3: Valid and invalid situations of the post-adaptation operation. “x” denotes that this bond can be broken manually.

have not yet done any bond-breaking on the product. However, **generated external bonds reveal the reaction site positions on the product, so we can manually break the bond by the legitimacy of reaction sites.** We term this process “post-adaption”, which serves the same functionality as the “reaction center prediction”, but with the information of reaction sites, it can be simplified as a rule-based process.

Specifically, traditional semi-template methods obey the following chemical principles:

- (a1) A broken bond is created in the product and the two product atoms corresponding to the broken bond are the reaction site;
- (a2) External groups will be strictly attached to the reaction site.

Our post-adaptation rule is exactly the inverse process of those chemical principles:

- (b1) Each generated external bond is attached to one of the atoms of the product. Based on (a2), this product atom must be a reaction site;
- (b2) Based on (a1), a broken bond produces two neighboring reaction sites. Thus, only if two reaction sites are attached to two neighboring product atoms the bond between these two atoms will be broken (the upper part of Figure 3), otherwise, it is invalid (the lower part of Figure 3).

Despite being a rule-based process, we find that it can obtain perfect identification performance. For more complex and extended scenarios, we can also refer to the implementation of “reaction center prediction” in previous work and train a predictive model to achieve it. This is the most scalable solution for general scenarios deserving to be explored in future work.

2.2 Prior Distribution and Interpolation Direction

Now we design the task-specific prior distribution (*i.e.* sampling start) P_{V_1} and P_{V_2} in the first two stages, and interpolation direction (*i.e.* transitional matrix) Q_t^X and Q_t^E . We denote n_g and n_x as the atom numbers of the external group \mathbf{g} and the product \mathbf{x} , respectively. In addition, we cannot predict the exact atom number of external groups in different cases, so we restrict n_g as a constant and create a dummy atom category. When denoising is complete, the atoms that are still in the dummy category will be deleted, and all remaining atoms constitute the real external group.

Prior Distribution. All atoms can start from a single absorbing distribution (Austin et al., 2021) v_x and all bonds can start from v_e . In the external group generation of stage 1, both atoms and bonds need to be denoised, but in the external bond generation of stage 2, only bonds need to be denoised. Therefore, the two prior distributions can be formulated as

$$P_{V_1} = p_{v_x}^{|n_g|} \times p_{v_e}^{|n_g|*|n_g|} \quad \text{and} \quad P_{V_2} = p_{v_e}^{|n_g|*|n_x|}. \quad (7)$$

For all atoms and bonds samples from the dummy state, we set the probability of single distribution as $p_{v_x} = [1, 0, 0, \dots, 0]^T \in \mathbb{R}^{1 \times (a+1)}$ and $p_{v_e} = [1, 0, 0, \dots, 0]^T \in \mathbb{R}^{1 \times (b+1)}$, where the first position in the vector denotes the dummy atom (or bond) category and the other positions denote each real categories (a types of atoms and b types of bonds), respectively.

Interpolation Direction. For the diffusion model to be reversible, any sample $s = (s_x, s_e) \sim p_{\text{data}}$ (p_{data} denotes the whole data distribution) must converge to a limit distribution q_∞ after t -step noise-applying, *i.e.*, $q_\infty = \lim_{t \rightarrow \infty} s Q_t$, which in turn is the sampling start. Therefore, we need to design Q_t^X and Q_t^E to satisfy that for any atom s_x and bond s_e from the data distribution, $p_{v_x} = \lim_{t \rightarrow \infty} s_x Q_t^X$ and $p_{v_e} = \lim_{t \rightarrow \infty} s_e Q_t^E$. Considering s_x and s_e are one-hot vectors, we compute $\lim_{t \rightarrow \infty} \bar{Q}_t^X = \mathbf{1}_x v_x^\top$ and $\lim_{t \rightarrow \infty} \bar{Q}_t^E = \mathbf{1}_e v_e^\top$, so a trivial design is

$$Q_t^X = \alpha_t \mathbf{I} + (1 - \alpha_t) \mathbf{1}_x v_x^\top, \quad Q_t^E = \alpha_t \mathbf{I} + (1 - \alpha_t) \mathbf{1}_e v_e^\top, \quad (8)$$

where \mathbf{I} is an identity matrix, $\mathbf{1}_x$ and $\mathbf{1}_e$ are all-one vectors, α_t is cosine schedule (Nichol & Dhariwal, 2021):

$$\bar{\alpha}_t = \cos\left(0.5\pi \frac{t/T + s}{1 + s}\right)^2. \quad (9)$$

2.3 Denoising Network for Training

We design $p_\theta(G_0|G_t, \mathbf{c})$ to model $p_\theta(G_{t-1}|G_t, \mathbf{c})$ at the above stages because the latter can be calculated

from the former according to Eq. 5. At time t , we merge the graph $\mathbf{G}_t = (\mathbf{X}_t, \mathbf{E}_t)$ and condition \mathbf{c} into a whole graph structure $\mathbf{G}_w = (\mathbf{G}_t, \mathbf{c})$, it is treated as the input of p_θ . Then we have the output $(p_{\mathbf{G}'_t}, p_{\mathbf{c}'}) = p_\theta(\mathbf{G}_w)$, where $p_{\mathbf{G}'_t} = (p_{\mathbf{X}'_t}, p_{\mathbf{E}'_t})$. The training loss of $p_\theta(\mathbf{G}_0|\mathbf{G}_t, \mathbf{c})$ is:

$$\begin{aligned} \mathcal{L} = & -\mu \cdot \sum_{x'_t \in \mathbf{X}'_t, x_0 \in \mathbf{X}_0} \text{cross-entropy}(p_{x'_t}, x_0) \\ & - \sum_{e'_t \in \mathbf{E}'_t, e_0 \in \mathbf{E}_0} \text{cross-entropy}(p_{e'_t}, e_0) \end{aligned} \quad (10)$$

where $\mathbf{G}_0 = (\mathbf{X}_0, \mathbf{E}_0)$ is the ground truth, and each x'_t - x_0 / e'_t - e_0 pair corresponds one-to-one in the graph position. μ is a control unit, specifically, in stage 1, $\mu = 0$, and in stage 2, μ is a hyperparameter to trade off the importance of atoms and bonds. In general, $\mu < 1$. We use the graph transformer architecture (Vignac et al., 2022; Yan et al., 2020) to design the network. Refer to Appendix B for network details.

3 Experiments

3.1 Setup

Dataset. We conduct experiments on the small-scale USPTO-50K (Schneider et al., 2016) and large-scale USPTO-full (Lowe, 2017) datasets. The former contains 50K single-step chemical reactions from 10 reaction types, and the latter consists of 760K training data that can demonstrate scalability. We follow standard splits (Schneider et al., 2016) to select 80%/10%/10% of data as training/validation/test sets.

Baseline. Our baselines can be divided into three groups: (i) *Template-based* methods, we choose RetroSim (Coley et al., 2017), NeuralSym (Segler & Waller, 2017b), GLN (Schneider et al., 2016), GraphRetro (Somnath et al., 2020), LocalRetro (Chen & Jung, 2021), RetroComposer (Yan et al., 2022), and RetroKNN (Xie et al., 2023). (ii) *Template-free* methods, we choose Transformer (Vaswani et al., 2017; Tetko et al., 2020), SCROP (Zheng et al., 2019), Tied Transformer (Kim et al., 2021), GTA (Seo et al., 2021), Graph2SMILES (Tu & Coley, 2022), Chemformer (Irwin et al., 2022), Retroformer (Wan et al., 2022), RootAligned (Zhong et al., 2022), RetroDCVAE (He et al., 2022), and RetroBridge (Igashov et al., 2023). (iii) *Semi-template* methods, we choose RetroXpert (Yan et al., 2020), G2G (Shi et al., 2020), RetroPrime (Wang et al., 2021) MEGAN (Sacha et al., 2021), and RootAligned (Zhong et al., 2022).

Implementation. We use open-source RDKit to construct molecular graphs based on molecular SMILES. For noise-applying and sampling processes, we set $T_1 = 500$ and $T_2 = 50$. For the training process, we train the graph transformer at 8-card 24G GTX-3090 with a training step of 100K, a batch size of 120,

and an Adam learning rate of 0.0001, and set $\mu = 0.2$. In addition, when setting n_g , to avoid extreme values that cause sparse distributions during the statistical process, we exclude all samples whose statistic is more than three times the standard deviation from the mean.

Evaluation. We follow prior works to adopt top- k accuracy as the main evaluation metric. For end-to-end models, beam search is adopted, but it is unfeasible for diffusion models. Therefore, we set the negative variational lower bound as the ranking score for each generated $\mathbf{G}_0 = (\mathbf{X}_0, \mathbf{E}_0)$:

$$\begin{aligned} \text{Score} = & \mu \cdot \mathbb{E}_{q(\mathbf{x}_0)} \mathbb{E}_{q(\mathbf{x}_t|\mathbf{x}_0)} [-\log p_\theta(x_0|x_t)] \\ & + \mathbb{E}_{q(\mathbf{e}_0)} \mathbb{E}_{q(\mathbf{e}_t|\mathbf{e}_0)} [-\log p_\theta(e_0|e_t)]. \end{aligned} \quad (11)$$

Note that this evaluation way strictly aligns with the evaluation in Igashov et al. (2023). For each sampling, the smaller the score is, the closer the sample is to the true data distribution. We sample 100 results for each case to rank the scores, and select the k lowest scoring results to compute top- k accuracy. In addition, we compute top- k validity that reflects the legitimacy of the reactants as chemical molecules. It is formulated as $\frac{1}{N \times k} \sum_1^N \sum_1^k \text{isvalid}(\mathbf{G}_0)$, where N denotes the dataset size. We also report **round-trip** accuracy and coverage (Schwaller et al., 2020) for all top- k samples.

3.2 Main Results

We report top- k accuracy and validity in the reactant class unknown setting and compare our method with all strong baselines. Specifically, we categorize our method as a *semi-template* method.

Accuracy. Table 1 shows the top- k accuracy results.

- On the USPTO-50k dataset, our method **outperforms all other competitive semi-template baselines** across different k values, particularly when $k = 5$. In addition, our method demonstrates competitive results when compared to the strongest template-free methods. Notably, our method holds a substantial advantage for $k > 1$.
- On the USPTO-full dataset, we compare our method with all methods whose original paper reported these results, because we cannot afford the extremely high cost of reproduction. Our method is also **the SOTA of the semi-template methods and outperforms all methods at $k = 1, 3, 5$** . In addition, our top-5 results outperform the top-10 results of most methods. These indicate that our method has a high potential for scalability.

In addition, we are surprised to find that the template-based methods exhibit extremely poor accuracies on the large-scale USPTO-full, which are the exact opposite performances on the small-scale USPTO-50k

Table 1: Top- k accuracy for the retrosynthesis task on USPTO-50K and USPTO-full dataset. RetroDiff achieves SOTA among *semi-template* methods. For *template-free* methods, RetroDiff is competitive, but the validity of generated molecules is significantly higher than theirs (Table 2); For *template-based* methods, RetroDiff has a slight performance disadvantage on USPTO-50k, but they perform much poorer than ours on large-scale USPTO-full, exhibiting poor scalability. Note: N/A indicates that the result was not reported in the original paper.

Method	Top- k accuracy in USPTO-50K				Top- k accuracy in USPTO-full			
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 1$	$k = 3$	$k = 5$	$k = 10$
Template-based methods								
RetroSim (Coley et al., 2017)	37.3	54.7	63.3	74.1	32.8	-	-	56.1
NeuralSym (Segler & Waller, 2017b)	44.4	65.3	72.4	78.9	35.8	-	-	60.8
GLN (Schneider et al., 2016)	52.5	69.0	75.6	83.7	39.0	50.1	55.3	61.3
GraphRetro (Somnath et al., 2020)	53.7	68.3	72.2	75.5			N/A	
LocalRetro (Chen & Jung, 2021)	53.4	77.5	85.9	92.4			N/A	
RetroComposer (Yan et al., 2022)	54.5	77.2	83.2	87.7	41.3	53.7	56.8	63.2
RetroKNN (Xie et al., 2023)	55.3	76.9	84.3	90.8			N/A	
Template-free methods								
Transformer (Vaswani et al., 2017)	42.4	58.6	63.8	67.7			N/A	
w/ Augmentation (<i>Aug.</i>) (Tetko et al., 2020)	48.3	-	73.4	77.4	44.4	-	-	73.3
SCROP (Zheng et al., 2019)	43.7	60.0	65.2	68.7			N/A	
Tied Transformer (Kim et al., 2021)	47.1	67.1	73.1	76.3			N/A	
GTA (Seo et al., 2021)	51.1	67.6	74.8	81.6	46.6	-	-	70.4
Graph2SMILES (Tu & Coley, 2022)	52.9	66.5	70.0	72.9	45.7	-	-	63.4
Chemformer (Irwin et al., 2022)	54.3	-	62.3	63.0			N/A	
Retroformer (Wan et al., 2022)	47.9	62.9	66.6	70.7			N/A	
w/ Augmentation (<i>Aug.</i>)	52.9	68.2	72.5	76.4			N/A	
RootAligned (Zhong et al., 2022)	44.0	67.5	74.0	76.2			N/A	
w/ 20 \times training <i>Aug.</i>	51.5	75.0	81.0	83.0			N/A	
w/ 20 \times training <i>Aug.</i> + 20 \times test <i>Aug.</i>	56.0	79.1	86.1	91.0			N/A	
RetroDCVAE (He et al., 2022)	53.1	68.1	71.6	74.3			N/A	
RetroBridge (Igashov et al., 2023)	50.8	74.1	80.6	85.6			N/A	
Semi-template methods								
RetroXpert (Yan et al., 2020)	50.4	61.1	62.3	63.4	N/A (Invalid Results ¹)			
G2G (Shi et al., 2020)	48.9	67.6	72.5	75.5			N/A	
RetroPrime (Wang et al., 2021)	51.4	70.8	74.0	76.1	44.1	59.1	62.8	68.5
MEGAN (Sacha et al., 2021)	48.1	70.7	78.4	86.1	33.6	-	-	63.9
RootAligned (Zhong et al., 2022)	49.1	68.4	75.8	82.2			N/A	
RetroDiff (ours)	52.6	71.2	81.0	85.3	46.9	60.4	65.1	70.3

Table 2: Top- k validity on USPTO-50K dataset of our method and *template-free* methods.

Model	Top- k validity			
	$k = 1$	$k = 3$	$k = 5$	$k = 10$
Transformer	97.2	97.9	82.4	73.1
Graph2SMILES	99.4	90.9	84.9	74.9
Retroformer (<i>Aug.</i>)	99.3	98.5	97.2	92.6
RetroDiff (ours)	99.2	99.0	97.8	94.3

dataset. This means that **compared with template-based methods, our methods possess greater scalability and are more adaptable to real-world large-scale application scenarios.**

Validity. Table 2 shows the top- k validity results. We compare our method with some strong template-free baselines whose original paper reported the result. We don’t compare template-based methods because they involve matching existing chemical templates, so they have few validity issues theoretically.

¹Data leakage leads to the invalid results. See <https://github.com/uta-smile/RetroXpert>.

We find that our validity score outperforms all template-free methods, especially as k increases. As for Retroformer, although it is a template-free method, it integrates the reaction center information defined in semi-template methods during the modeling process. This further reflects that the prior of semi-template methods bring greater validity improvement. Overall, **compared with template-free methods, our method has a great advantage in generating validity, which reduces unavailability and security risks in practical applications.**

Round-trip Metrics. For the top- k samples generated for each product, we assess round-trip accuracy and coverage using the Molecular Transformer model (Schwaller et al., 2019). Round-trip accuracy reflects the percentage of correctly predicted reactants out of all predictions, where a reactant is deemed correct if it matches the ground truth or successfully regenerates the original product. Round-trip coverage indicates whether at least one correct prediction appears in the top- k samples. These metrics highlight that a single product can correspond to multiple valid sets of reactants. Table 3 displays the results on the

Table 3: Top- k round-trip results on USPTO-50k.

Method	Top- k Coverage			Top- k Accuracy		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
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
MEGAN	78.1	88.6	91.3	78.1	67.3	61.7
Graph2SMILES	-	-	-	76.7	56.0	46.4
Retroformer	-	-	-	78.6	71.8	67.1
RetroBridge	85.1	95.7	97.1	85.1	73.6	67.8
RetroDiff (ours)	86.3	96.2	97.6	84.5	75.3	69.2

 Table 4: Top- k accuracy of “external group generation” sub-module (* indicates the performance of raw synthon completion).

Model	Top- k accuracy			
	$k = 1$	$k = 3$	$k = 5$	$k = 10$
G2G*	61.1	81.5	86.7	90.0
RetroXpert*	64.8	77.6	80.8	84.5
RetroDiff (ours)	66.5	78.4	85.0	86.4

USPTO-50k dataset, showing that RetroDiff maintains strong round-trip performance and achieves state-of-the-art coverage for all k values.

3.3 Ablation

In this part, We conduct ablation studies to analyze the sub-module performances in each stage, *i.e.*, external group generation and external bond generation.

External Group Generation. RetroDiff first generates external groups given raw products. In traditional semi-template methods, the external group generation equates to the synthon completion, commonly addressed in two distinct ways: (i) auto-regressive generation, including encoder-decoder sequence prediction (Shi et al., 2020) and action-state sequence prediction (Somnath et al., 2020), (ii) finite-space search, where all possible leaving group vocabularies are constructed in a database, followed by maximum likelihood estimation using a classifier (Yan et al., 2020). In our setting, the external group generation is treated as non-autoregressive generation.

Table 4 shows the results and we compare the external group generation performance between RetroDiff and the synthon completion performance of other methods. Our external group generation outperforms the rest of the methods on top-1, but not as well as G2G when $k > 1$, albeit within a reasonable margin. A plausible explanation lies in the fact that G2G acquires information about the reaction center when generating the external group, *i.e.*, serial complementation from the reaction sites. In contrast, RetroDiff lacks this specific information, resulting in a slight disadvantage.

 Table 5: Top- k accuracy of “external bond generation” sub-module (* indicates the performance of raw reaction center prediction).

Model	Top- k accuracy			
	$k = 1$	$k = 3$	$k = 5$	$k = 10$
G2G*	61.1	81.5	86.7	90.0
RetroXpert*	64.3	-	-	-
GraphRetro*	75.6	87.4	92.5	96.1
RetroDiff (ours)	82.3	92.4	95.5	96.8

External Bond Generation. Next, RetroDiff generates external bonds given products and generated external groups. In the traditional semi-template methods, reaction centers are predicted directly by the classification model, whereas under our template setup, this task equates to a combination of external bond generation and post-adaptation. Thus, we conduct a direct comparison between the performance of previous methods in predicting reaction centers and the external bond generation performance of our model. Table 5 shows the results, indicating that predicting the connecting bond between the product and the external group, and thus deducing the reaction center based on the rule, can achieve higher accuracy than the direct prediction of the reaction center given the product.

Specifically, the atom number of the product is denoted as n , the bond number as m , and the external group atom number as r . Considering a maximum bonding site limit of 4 for an atom (*e.g.* Carbon atom) excluding Hydrogen atoms, we establish the condition $m \leq 2n$. In the realm of traditional reaction center prediction, the search space size is m , whereas, for external bond generation, it is rn . Consequently, the complexity of the external bond generation task is higher than that of the reaction center prediction task. However, the external bond generation task leverages molecular information from external groups, expanding the model’s ability to search for reaction sites more accurately by incorporating additional chemical insights. Consequently, the observed superior performance of external bond generation over traditional reaction center prediction can be empirically attributed to the enriched chemical information acquired through the former.

4 Related Work and Discussion

Related Work. Our research focuses on the important biochemical topic of molecular retrosynthesis, which can be categorized into three types: template-based, template-free, and semi-template methods. We summarize the existing retrosynthesis studies in Appendix A.1. Additionally, our research methodology

involves diffusion models, which have promising applications in the field of molecular generation. Therefore, we introduce existing applications of diffusion models in Appendix A.2.

Discussion about RetroDiff. We also conduct some interesting discussions about RetroDiff in the appendix due to space limits: (I) Why select the absorbing distribution as the prior (Appendix C.1); (II) Why serial multi-stage modeling is used instead of single-stage joint modeling (Appendix C.2); (III) Performances on all reaction types (Appendix C.3); (IV) Generation-process visualizations (Appendix C.4).

5 Conclusion

We introduce RetroDiff, a multi-stage conditional retrosynthesis diffusion model. Considering maximizing the usage of chemical information in the molecule, we reset the template to decompose the retrosynthesis into external group generation and external bond generation sub-tasks, and set a joint diffusion model to transfer dummy distributions to group and bond distributions serially. Our method performs the best under the semi-template setting in the accuracy and validity evaluation metrics. In the future, we will try to extend our RetroDiff to multi-step retrosynthesis scenarios.

Acknowledgements

This work is supported by the the Natural Science Foundation of China (Grant No. 62376133), Beijing Nova Program (20240484682) and Wuxi Research Institute of Applied Technologies, Tsinghua University under Grant 20242001120.

References

- Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.
- Benson Chen, Tianxiao Shen, Tommi S Jaakkola, and Regina Barzilay. Learning to make generalizable and diverse predictions for retrosynthesis. *arXiv preprint arXiv:1910.09688*, 2019.
- Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.
- Connor W Coley, Luke Rogers, William H Green, and Klavs F Jensen. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science*, 3(12):1237–1245, 2017.
- Elias James Corey. *The logic of chemical synthesis*. 1991.
- Elias James Corey and W Todd Wipke. Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969.
- Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2022.
- Hua-Rui He, Jie Wang, Yunfei Liu, and Feng Wu. Modeling diverse chemical reactions for single-step retrosynthesis via discrete latent variables. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 717–726, 2022.
- Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. Sequence to sequence mixture model for diverse machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 583–592, 2018.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Ilia Igashov, Hannes Stärk, Clement Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael M Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022.
- Ilia Igashov, Arne Schneuing, Marwin Segler, Michael Bronstein, and Bruno Correia. Retrobridge: Modeling retrosynthesis with markov bridges. *arXiv preprint arXiv:2308.16212*, 2023.
- Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022.
- Shaojie Jiang and Maarten de Rijke. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pp. 81–86, 2018.
- Jaehyeong Jo, Seul Lee, and Sung Ju Hwang. Score-based generative modeling of graphs via the system of stochastic differential equations. In *International Conference on Machine Learning*, pp. 10362–10383. PMLR, 2022.

- Eunji Kim, Dongseon Lee, Youngchun Kwon, Min Sik Park, and Youn-Suk Choi. Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *Journal of Chemical Information and Modeling*, 61(1):123–133, 2021.
- Daniel Lowe. Chemical reactions from us patents (1976-sep2016). 2017.
- Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.
- Nicholas Roberts, Davis Liang, Graham Neubig, and Zachary C Lipton. Decoding and diversity in machine translation. *arXiv preprint arXiv:2011.13477*, 2020.
- Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.
- Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. What’s what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346, 2016.
- Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.
- Philippe Schwaller, Riccardo Petraglia, Valerio Zullo, Vishnu H Nair, Rico Andreas Haeuselmann, Riccardo Pisoni, Costas Bekas, Anna Iuliano, and Teodoro Laino. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science*, 11(12):3316–3325, 2020.
- Marwin HS Segler and Mark P Waller. Modelling chemical reasoning to predict and invent reactions. *Chemistry—A European Journal*, 23(25):6118–6128, 2017a.
- Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry—A European Journal*, 23(25):5966–5971, 2017b.
- Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018.
- Seung-Woo Seo, You Young Song, June Yong Yang, Seohui Bae, Hankook Lee, Jinwoo Shin, Sung Ju Hwang, and Eunho Yang. Gta: Graph truncated attention for retrosynthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 531–539, 2021.
- Chence Shi, Minkai Xu, Hongyu Guo, Ming Zhang, and Jian Tang. A graph to graphs framework for retrosynthesis prediction. In *International conference on machine learning*, pp. 8818–8827. PMLR, 2020.
- Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.
- Vignesh Ram Somnath, Charlotte Bunne, Connor W Coley, Andreas Krause, and Regina Barzilay. Learning graph models for template-free retrosynthesis. *arXiv preprint arXiv:2006.07038*, 2020.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):5575, 2020.
- Zhengkai Tu and Connor W Coley. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling*, 62(15):3503–3513, 2022.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. In *The Eleventh International Conference on Learning Representations*, 2022.
- Yue Wan, Chang-Yu Hsieh, Ben Liao, and Shengyu Zhang. Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In *International Conference on Machine Learning*, pp. 22475–22490. PMLR, 2022.
- Xiaorui Wang, Yuquan Li, Jiezhong Qiu, Guangyong Chen, Huanxiang Liu, Benben Liao, Chang-Yu Hsieh, and Xiaojun Yao. Retroprime: A diverse, plausible and transformer-based method

for single-step retrosynthesis predictions. *Chemical Engineering Journal*, 420:129845, 2021.

Shufang Xie, Rui Yan, Junliang Guo, Yingce Xia, Lijun Wu, and Tao Qin. Retrosynthesis prediction with local template retrieval. *arXiv preprint arXiv:2306.04123*, 2023.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021.

Chaochao Yan, Qianggang Ding, Peilin Zhao, Shuangjia Zheng, Jinyu Yang, Yang Yu, and Junzhou Huang. Retroxpert: Decompose retrosynthesis prediction like a chemist. *Advances in Neural Information Processing Systems*, 33:11248–11258, 2020.

Chaochao Yan, Peilin Zhao, Chan Lu, Yang Yu, and Junzhou Huang. Retrocomposer: composing templates for template-based retrosynthesis prediction. *Biomolecules*, 12(9):1325, 2022.

Shuangjia Zheng, Jiahua Rao, Zhongyue Zhang, Jun Xu, and Yuedong Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of chemical information and modeling*, 60(1):47–55, 2019.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Min Wu, Tingjun Hou, and Mingli Song. Root-aligned smiles: a tight representation for chemical reaction prediction. *Chemical Science*, 13(31):9023–9034, 2022.

Checklist

- For all models and algorithms presented, check if you include:
 - A clear description of the mathematical setting, assumptions, algorithm, and/or model. [\[Yes\]](#)
 - An analysis of the properties and complexity (time, space, sample size) of any algorithm. [\[Not Applicable\]](#)
 - (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
- For any theoretical claim, check if you include:
 - Statements of the full set of assumptions of all theoretical results. [\[Not Applicable\]](#)
 - Complete proofs of all theoretical results. [\[Not Applicable\]](#)
 - Clear explanations of any assumptions. [\[Not Applicable\]](#)
- For all figures and tables that present empirical results, check if you include:
 - The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [\[Yes\]](#)
 - All the training details (e.g., data splits, hyperparameters, how they were chosen). [\[Yes\]](#)
 - A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [\[Not Applicable\]](#)
 - A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [\[Yes\]](#)
- If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
 - Citations of the creator If your work uses existing assets. [\[Yes\]](#)
 - The license information of the assets, if applicable. [\[Not Applicable\]](#)
 - New assets either in the supplemental material or as a URL, if applicable. [\[Not Applicable\]](#)
 - Information about consent from data providers/curators. [\[Yes\]](#)
 - Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [\[Not Applicable\]](#)
- If you used crowdsourcing or conducted research with human subjects, check if you include:
 - The full text of instructions given to participants and screenshots. [\[Not Applicable\]](#)
 - Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [\[Not Applicable\]](#)
 - The estimated hourly wage paid to participants and the total amount spent on participant compensation. [\[Not Applicable\]](#)

RetroDiff: Retrosynthesis as Multi-stage Distribution Interpolation

Supplementary Materials

A Related Work

A.1 Retrosynthesis Prediction

Existing methods of retrosynthesis prediction can be broadly categorized into three groups: (i) *Template-based* methods retrieve the best match reaction template for a target molecule from a large-scale chemical database, they focus on computing the similarity scores between target molecules and templates using either plain rules (Coley et al., 2017) or neural networks (Schneider et al., 2016; Somnath et al., 2020; Chen & Jung, 2021). (ii) *Template-free* methods adopt end-to-end generative models to directly obtain final reactants given products (Zheng et al., 2019; Kim et al., 2021; Seo et al., 2021; Tu & Coley, 2022; Wan et al., 2022). Despite the efficiencies of data-driven methods, the chemical prior has been ignored. (iii) *Semi-template* methods combine the advantages of the above two approaches, they split the task into two parts, *i.e.*, reaction center prediction and synthon correction (Yan et al., 2020; Shi et al., 2020; Wang et al., 2021), followed by serial modeling using a classification model and a generative model, respectively.

A.2 Diffusion Models in Molecules

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) is a class of score-based generative models (Song & Ermon, 2019), whose goal is to learn the latent structure of a dataset by modeling how data points diffuse through the latent space. Since the generalized discrete diffusion model (Austin et al., 2021) and the discrete graph diffusion model (Vignac et al., 2022) have been proposed, the molecular design field began to use them extensively, such as molecular conformation (Xu et al., 2021), molecular docking (Corso et al., 2022), and molecular linking (Igashov et al., 2022). To our knowledge, we are the first to apply discrete diffusion models to the retrosynthesis prediction task. (Igashov et al., 2023) have done similar work using a diffusion model, and they achieve the template-free retrosynthesis prediction. However, our method performs better and has stronger chemical interpretability.

B Denoising Network for Training

B.1 Network Architecture

The overall architecture and the graph transformer module for each layer are shown in Figure 4. Specifically, we add the global feature \mathbf{f} to the input, so the final input at time t is $(\mathbf{G}_w, \mathbf{f}) = (\mathbf{X}, \mathbf{E}, \mathbf{f})$. For the global feature \mathbf{f} , we obtain the topological features and chemical features (Details can be seen in Appendix B.2) of this molecular graph to splice with the original features. After the pre-processing, $\mathbf{G} = (\mathbf{X}, \mathbf{E}, \mathbf{f})$ is input to a feed-forward network to be encoded, then it will pass serially through the n_{layer} graph transformer modules. Finally, another feed-forward network is set to decode the graph features, the output is the final prediction result.

B.2 Additional Input Features

To fully explore the potential features of a molecular graph, we can analyze it from two perspectives: topological features and chemical features following (Vignac et al., 2022; Yan et al., 2020).

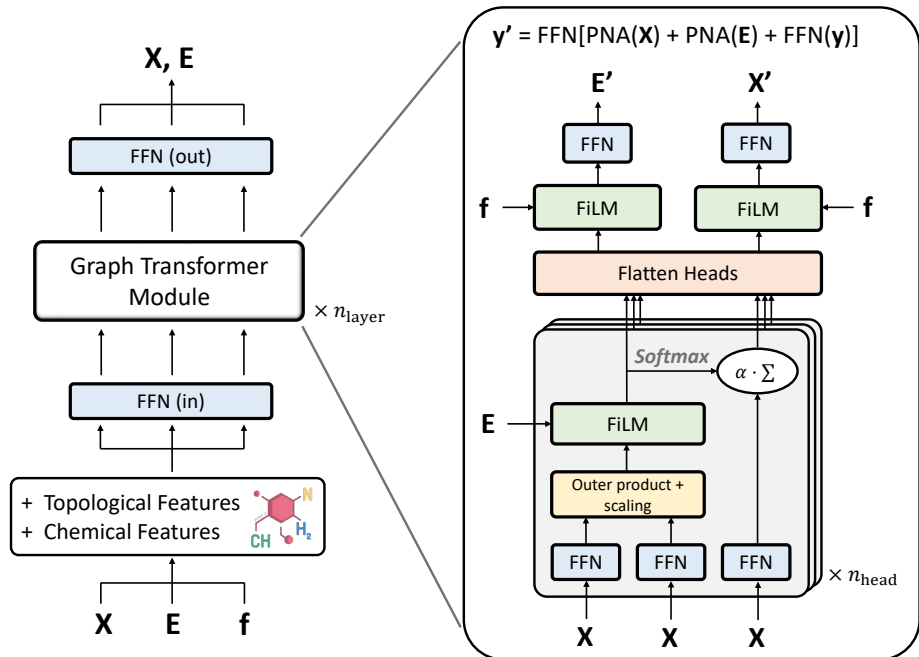


Figure 4: The whole architecture (left) of the denoising network for training with graph transformer modules (right). \mathbf{X} , \mathbf{E} , \mathbf{y} denote the atom features, bond features, and global features, respectively. $\text{FiLM}(\mathbf{M}_1, \mathbf{M}_2) = \mathbf{M}_1 \mathbf{W}_1 + (\mathbf{M}_2 \mathbf{W}_2) \odot \mathbf{M}_2 + \mathbf{M}_2$, where $\mathbf{W}_1, \mathbf{W}_2$ are learnable. $\text{PNA}(\mathbf{M}) = [\max(\mathbf{M}) \odot \min(\mathbf{M}) \odot \text{mean}(\mathbf{M}) \odot \text{std}(\mathbf{M})] \mathbf{W}$, where \mathbf{W} is learnable.

Topological Features. We focus on two useful topological features. First is the **spectral features**, we first compute some graph-level features that relate to the eigenvalues of the graph Laplacian: the number of connected components (given by the multiplicity of eigenvalue 0), as well as the 5 first nonzero eigenvalues. We then add node-level features relative to the graph eigenvectors: an estimation of the biggest connected component (using the eigenvectors associated with eigenvalue 0), as well as the two first eigenvectors associated with non-zero eigenvalues.

Second is the **cycle detection**. To further refine it, we split it into node-level and graph-level features. For node-level features, we compute how many k -cycles this node belongs to, where $3 \leq k \leq 5$. The feature formulas are as follows:

$$\begin{aligned} \mathbf{X}_3 &= \text{diag}(\mathbf{A}^3)/2, \\ \mathbf{X}_4 &= (\text{diag}(\mathbf{A}^4) - \mathbf{d}(\mathbf{d} - 1) - \mathbf{A}(\mathbf{d}\mathbf{1}_n^\top)\mathbf{1}_n)/2, \\ \mathbf{X}_5 &= (\text{diag}(\mathbf{A}^5) - 2\text{diag}(\mathbf{A}^3) \odot \mathbf{d} - \mathbf{A}(\text{diag}(\mathbf{A}^3)\mathbf{1}_n^\top)\mathbf{1}_n + \text{diag}(\mathbf{A}^3))/2, \end{aligned} \quad (12)$$

where \mathbf{d} denotes the vector containing node degrees. For graph-level features, we compute how many k -cycles this graph contains, where $3 \leq k \leq 6$. The feature formulas are as follows:

$$\begin{aligned} \mathbf{y}_3 &= \mathbf{X}_3^\top \mathbf{1}_n / 3, \\ \mathbf{y}_4 &= \mathbf{X}_4^\top \mathbf{1}_n / 4, \\ \mathbf{y}_5 &= \mathbf{X}_5^\top \mathbf{1}_n / 5, \\ \mathbf{y}_6 &= \text{Tr}(\mathbf{A}^6) - 3\text{Tr}(\mathbf{A}^3 \odot \mathbf{A}^3) + 9\|\mathbf{A}(\mathbf{A}^2 \odot \mathbf{A}^2)\|_F - 6\text{diag}(\mathbf{A}^2)^\top \text{diag}(\mathbf{A}^4) \\ &\quad + 6\text{Tr}(\mathbf{A}^4) - 4\text{Tr}(\mathbf{A}^3) + 4\text{Tr}(\mathbf{A}^2 \dot{\mathbf{A}}^2 \odot \mathbf{A}^2) + 3\|\mathbf{A}^3\|_F - 12(\mathbf{A}^2 \odot \mathbf{A}^2) + 4\text{Tr}(\mathbf{A}^2), \end{aligned} \quad (13)$$

where $\|\cdot\|_F$ is Frobenius norm.

Chemical Features. There are two useful chemical features. First is the **atom valency**, which can be concatenated to the atom features \mathbf{X} . Second is the **molecular weight**, which can be concatenated to the global

features y .

C Discussion: Extended Analysis

C.1 Prior Distribution Selection

In our setting of the prior distribution, we choose the absorbing distribution (Austin et al., 2021), which is a special kind of marginal distribution (Vignac et al., 2022) from a generalized perspective, *i.e.* collapses the probabilities at all positions to those at one of them, making the initial state practically deterministic. We use it because the noisy graph we define is a dummy graph, which means that the probability of the empty category is 1 and others 0.

Another common prior distribution is the uniform distribution, meaning the sampling starts from a completely random state. We compare the two distributions in the whole process, the external group generation process, and the external bond generation process in Table 6.

Table 6: Top- k accuracy under different prior distributions.

Prior Dist.	Top- k accuracy											
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 1$	$k = 3$	$k = 5$	$k = 10$
	Overall				External Group Generation				External Bond Generation			
Uniform	51.7	70.1	79.6	83.2	66.2	78.8	84.6	86.2	80.2	89.8	91.7	92.9
Absorbing	52.6	71.2	81.0	85.3	66.5	78.4	85.0	86.4	82.3	92.4	95.5	96.8

We find that absorbing distribution performs slightly better than uniform distribution, and the difference mainly appears in the external bond generation. Since most of the external bonds are EMPTY in the ground truth, which means that the bond categories of most positions do not change during the diffusion process, setting an absolute distribution distribution helps faster convergence and accurate learning. From this perspective, we find that a low-entropy marginal prior distribution (*e.g.* absorbing distribution) is more suitable for predictive tasks like external bond generation.

C.2 Single-stage Diffusion *vs.* Multi-stage Diffusion

There are three modes of the modeling approach: (1) joint modeling, where the group and the bond are generated at the same time; (2) generating the group first and then the bond; (3) generating the bond first and then the group. The first is the single-stage diffusion, and the latter two are the multi-stage diffusions. We compare the performances of the three modes, as shown in Table 7.

Table 7: Top- k accuracy under different diffusion modeling modes.

Modeling Mode	Top- k accuracy			
	$k = 1$	$k = 3$	$k = 5$	$k = 10$
Joint Modeling	51.3	69.6	79.8	84.1
First Group, Then Bond	52.6	71.2	81.0	85.3
First Bond, Then Group	49.8	66.1	76.7	81.4

Previous work (Jo et al., 2022) concluded that joint modeling has better performance than “marginal then conditional” serial modeling. However, in our initial explorations, we find that mode (2), the serial mode of generating groups first and then bonds, was more effective for the retrosynthesis task. We speculate that the underlying reason is that **the information discrepancy of groups and bonds is large**. The information of additional bond information might not affect the group generation too much, whereas additional group information might largely determine the formation of a bond. Thus, generating groups first brings a large positive effect on

Table 8: Top- k accuracy on all reaction classes.

Reaction Class	Reaction Fraction(%)	Top- k accuracy in USPTO-50K			
		$k = 1$	$k = 3$	$k = 5$	$k = 10$
heteroatom alkylation and arylation	30.3	51.4	68.4	80.0	84.6
acylation and related processes	23.8	60.2	78.0	87.6	90.2
deprotections	16.5	49.1	75.7	82.4	88.6
C-C bond formation	11.3	41.7	60.3	71.2	75.3
reductions	9.2	58.9	76.8	82.5	88.2
functional group interconversion	3.7	30.5	51.0	62.7	68.4
heterocycle formation	1.8	47.2	68.3	76.4	78.2
oxidation	1.6	73.6	83.4	90.8	91.7
protections	1.3	72.1	87.8	88.2	89.8
functional group addition	0.5	84.0	84.0	86.3	88.0

bond generation. Therefore, we adopt the multi-stage diffusion in our work. This is an interesting phenomenon whose underlying reasons deserve to be explored in future work.

C.3 Performance in All Reaction Types

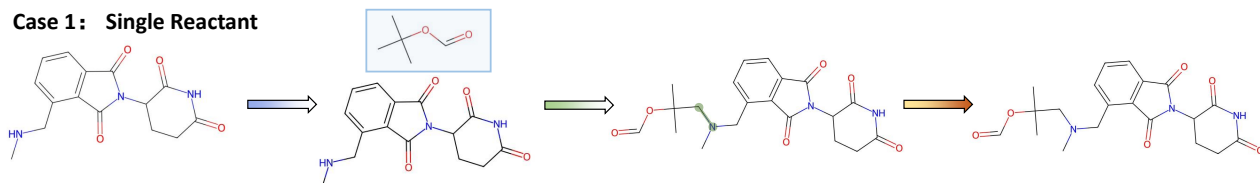
We list all reaction classes on the USPTO-50k dataset, and report the top- k accuracy of each reaction class when trained with reaction class unknown in Table 8. From the results, we have the following analyses:

- For some reactions, such as functional group addition, oxidations, and the protections, the accuracy is significantly higher than the average.
- For some other reactions, such as functional group interconversion and C-C bond formation, the accuracy is significantly lower than the average.
- These observation helps us better understand RetroDiff’s strengths and limitations on different reactions, improving the interpretability of RetroDiff.

C.4 Case Study via Visualization

In this part, we present visualizations of both successful and failed cases to provide an intuitive analysis of RetroDiff’s mechanisms. Figure 5 illustrates instances of success, featuring external groups delineated by blue shaded boxes and external bonds highlighted in green. Conversely, Figure 6 showcases failed cases, revealing two prevalent situations associated with higher error rates: (i) elevated error rates are observed when the external group size is substantial, leading to biases in the prediction of bonds between atoms, and (ii) for external bond predictions, inaccuracies in predicting reaction sites on the product contribute to ineffective post-adaptation of reaction centers.

Case 1: Single Reactant



Case 2: Double Reactants

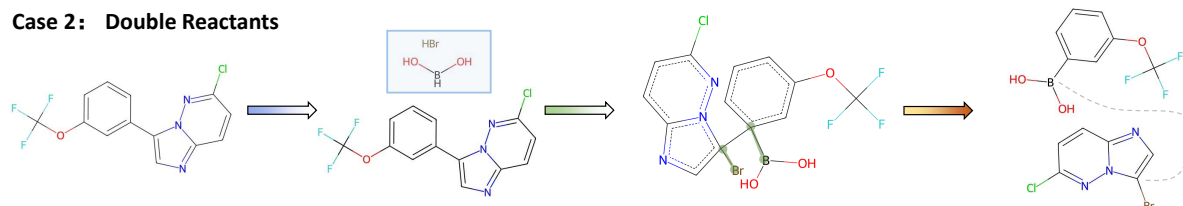
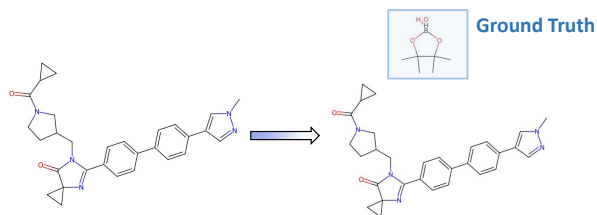
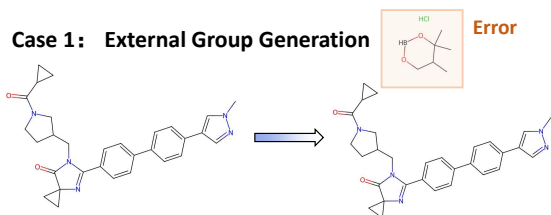


Figure 5: Successful cases produced by RetroDiff on the retrosynthesis task.

Case 1: External Group Generation



Case 2: External Bond Generation

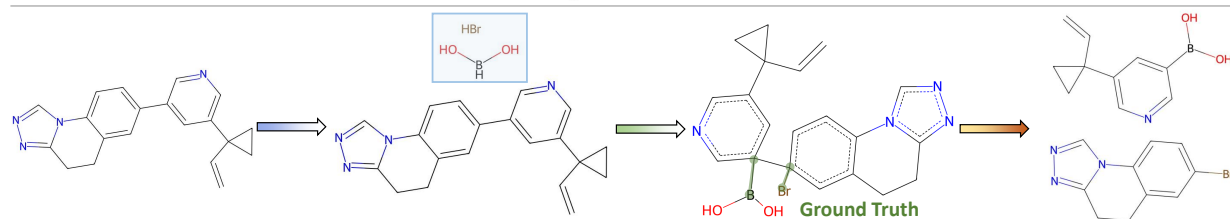
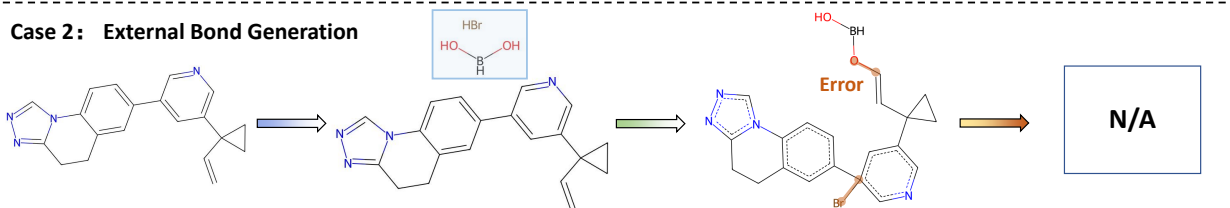


Figure 6: Failed cases produced by RetroDiff on the retrosynthesis task.