

DIFFER²: DIFFUSION ENSEMBLES FOR RETROSYNTHESIS PREDICTION WITH SMILES ADAPTED PARTICLE GUIDANCE

Anonymous authors

Paper under double-blind review

ABSTRACT

Computational retrosynthesis prediction has the potential to reduce development time for newly discovered drugs by automatically generating potential reactions for a target product. Data-driven approaches commonly treat this as a sequence-to-sequence generation task on SMILES strings. In this work, we construct ensembles of discrete-time and continuous-time diffusion models for molecular generation and incorporate guidance mechanisms for improved output diversity. We propose an adaptation of particle guidance to SMILES sequence generation which significantly improves the number of unique molecules generated by diffusion ensembles while increasing top-k accuracy. These results further expand the efficacy of discrete diffusion for SMILES generation, and our empirical analyses offer new insights into the capabilities of diffusion models for chemical retrosynthesis.

1 INTRODUCTION

The goal of retrosynthesis is to construct reactant molecules that can be used to synthesize a target product. Retrosynthesis prediction constitutes a pivotal step in the drug manufacturing pipeline. Repeatedly applying single-step retrosynthesis models to a given problem can result in full synthetic pathway discovery (Segler et al., 2018; Shen et al., 2021), further supporting pharmaceutical advancement for new drug modalities. However, even single-step retrosynthesis prediction is difficult, in part due to the diversity of reactant sets that might map to a single product. While forward synthesis prediction is primarily considered a surjective process with a deterministic output, retrosynthesis prediction is in many ways the opposite, with many possible outputs resulting from a single input. As a result, researchers have taken a particular interest in deep learning methods that hasten the process of retrosynthesis discovery by quickly and efficiently predicting multiple possible reactant sets from a given product molecule.

Numerous methodologies have been developed, from template-based methods that use mined reaction mechanisms to predict reactants to template-free approaches that directly predict string or graph representations of reactant molecules conditioned on the target product. The most successful of these template-free approaches use traditional transformer encoder-decoder architectures on SMILES (Weininger, 1988) string representations of the molecules, effectively treating the problem as a machine translation task (Zhong et al., 2022). A recent approach, DiffER (Current et al., 2024), explored the usage of discrete diffusion models for the retrosynthesis task and provided evidence that an ensemble of discrete-time diffusion models could be a practical approach for sampling from the posterior distribution of reactants. DiffER outperformed the state-of-the-art R-SMILES (Zhong et al., 2022) at top-1 accuracy on the commonly used USPTO-50k dataset (Lowe, 2017), but suffered from low output diversity due to the nature of posterior sampling.

In this work, we extend upon the work of DiffER (Current et al., 2024) by adapting their approach to include continuous-time discrete diffusion models and applying classifier-based particle guidance (Corso et al., 2023) to SMILES string generation. Both of these changes aim to improve the efficacy and output diversity of models over the original DiffER. A visual overview of the DiffER² ensemble approach is presented in Figure 1. The main improvements on the original DiffER methodology are as follows:

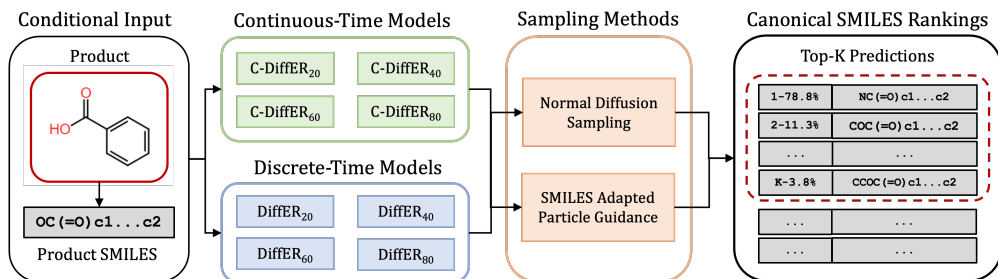


Figure 1: Overview of the DifferER² ensemble. The DifferER² ensemble contains both continuous-time (C-DifferER) and discrete-time (DifferER) models to leverage differing capabilities between the two diffusion archetypes. Both normal sampling methods and SMILES adapted particle guidance are used to generate output. Output SMILES are ranked according to frequency of the canonical form. DifferER _{κ} indicates the value of κ as described in Section 3.2.

1. We incorporate continuous-time discrete diffusion models into the DifferER ensemble to leverage different capabilities of discrete-time and continuous-time diffusion.
2. We explore the efficacy of SMILES-adapted guidance mechanisms to further enhance the diversity and accuracy of diffusion models for chemical retrosynthesis.
3. Finally, we offer an in-depth exploration into the differing performance capabilities of auto-regressive and discrete-space diffusion models on the chemical retrosynthesis task.

By combining both continuous-time and discrete-time diffusion models with classifier-based particle guidance (Corso et al., 2023), we significantly improve ensemble performance and reduce oversampling of posterior peaks while maintaining or boosting top-k accuracy across the board. Our results enhance the efficacy of diffusion ensembles for SMILES generation while providing an empirical analysis of diffusion performance compared to auto-regressive approaches for sequence generation.

2 RELATED WORK

2.1 RETROSYNTHESIS PREDICTION

Data-driven models for retrosynthesis prediction are commonly divided into three approaches: template-based, semi-template, and template-free methods. Common template-based approaches include GLN (Dai et al., 2019) and LocalRetro (Chen & Jung, 2021). These approaches primarily focus on identifying locations in a molecule where reaction templates can be applied and ranking the efficacy and feasibility of those templates on the target area. Semi-template methods include the likes of GraphRetro (Somnath et al., 2021), G²Retro (Chen et al., 2023), and GDiffRetro (Sun et al., 2025). These approaches generally work in a two-step manner: in the first step, a reaction center is first predicted from the product molecule, and the molecule is subsequently split into synthons. In the second step, each synthon is completed to construct a final reactant molecule. Notably, GDiffRetro (Sun et al., 2025) uses discrete-time diffusion processes to complete predicted synthons. Finally, template-free methods such as Chemformer (Irwin et al., 2022), MEGAN (Sacha et al., 2021), R-Smiles (Zhong et al., 2022), and DifferER (Current et al., 2024) directly construct reactant molecules from the product molecule. This is most often done in a sequence-to-sequence manner using SMILES strings (Irwin et al., 2022; Zhong et al., 2022; Current et al., 2024), though some approaches directly operate on graphs, such as the case of MEGAN (Sacha et al., 2021), which iteratively builds the reactant molecule(s) through a series of edits. DifferER (Current et al., 2024) is the first approach to use conditional discrete diffusion to accomplish the retrosynthesis prediction task, although their methodology suffers from reduced output diversity and high computational cost compared to auto-regressive decoding strategies. We primarily focus on template-free methods for performance comparison.

2.2 DISCRETE-SPACE DIFFUSION

Generative diffusion models have been a significant topic of research in recent years, and have achieved significant success in the realms of image generation (Croitoru et al., 2023; Yang et al., 2023) and signal processing (Yang et al., 2023). Though diffusion models traditionally operate in the continuous regime (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Song et al., 2020; Ho et al., 2020; Saharia et al., 2022; Gu et al., 2022; Zhang et al., 2023), significant effort has been made extending diffusion models to discrete spaces (Hooeboom et al., 2021; Austin et al., 2021; Gong et al., 2022; 2023; He et al., 2022; Yuan et al., 2022; Dieleman et al., 2022; Ghazvininejad et al., 2019). Discrete-space models have demonstrated remarkable performance for molecular generation (Xu et al., 2022; Alakhdar et al., 2024; Hooeboom et al., 2022; Xu et al., 2023; Schneuing et al., 2024) and protein design (Gruver et al., 2023; Watson et al., 2022; Ni et al., 2023), but have struggled to out-compete their auto-regressive counterparts on discrete sequence generation. Furthermore, most work in this space has primarily focused on discrete-time diffusion. Only recently has discrete-space diffusion been extended to continuous-time implementations (Campbell et al., 2022; Zhao et al., 2024; Xu et al., 2024), which demonstrate improved modeling capabilities and sampling efficiency compared to their discrete-time parallels. Furthermore, incorporating recent advances in guidance mechanisms during diffusion sampling (Bansal et al., 2023; Liu et al., 2023; Epstein et al., 2023; Ho & Salimans, 2022) remains a challenge, due to the difficulty of applying gradient-based approaches on a discrete domain (Schiff et al., 2024; Gruver et al., 2023). Schiff et al. (2024) recently adapted classifier-free and classifier-based guidance mechanisms for discrete diffusion and demonstrated promising results on molecular generation tasks, though more work needs to be done constructing and evaluating guidance mechanisms for discrete diffusion tasks.

3 METHODS

Following the original DiffER (Current et al., 2024), we treat retrosynthesis prediction as a discrete sequence-to-sequence generation task, where the target product is a conditional input to the diffusion generation akin to text-to-image models. We employ both discrete-time and continuous-time diffusion models. Like DiffER, we incorporate length prediction strategies to improve diffusion sampling as well as enable robust model ensembling. Finally, we adapt classifier-based particle guidance (Corso et al., 2023) to SMILES sequence generation to improve output diversity and reduce over-sampling of posterior peaks. By applying our techniques to SMILES generation, we are able to offer direct comparisons to prior works, and more clearly demonstrate the effectiveness of our diffusion ensembling and guidance methodologies. However, we recognize the existence of other molecular string representations such as SELFIES (Krenn et al., 2019). We offer a comparison to a SEFLIES-based implementation of our work in Appendix A.4.

3.1 PRELIMINARIES

Following the notation of Zhao et al. (2024), let $x_0 \sim p_{data}(x_0)$ be a categorical random variable representing the observed data with discrete distribution $p_{data}(x_0; K)$. Let $x_t \sim q(x_t)$ be the latent variable at time $t \in [0, 1]$ (continuous-time, max time $T = 1$) or $t \in \{1, \dots, T\}$ (discrete-time) and $x_{t|s} \sim q(x_t|x_s)$ be the conditional random variable. Let $x_t^{1:L_x} \in \{0, 1\}^{L_x \times K}$ be a sequence of length L_x at time t , and let $x_t^{[\ell]}$ be the sequence indexed at position ℓ . We assume each discrete random variable is governed by a categorical distribution $\mathcal{C}$ with probability distribution $p \in [0, 1]^K$, where $\|p\|_1 = 1$, such that $x_t \sim \mathcal{C}(x_t; p)$. Furthermore, let $Q_{t|s} \in [0, 1]^{K \times K}$ represent the transition matrix between times s and t , such that $[Q_{t|s}]_{ij} = q(x_t = e_i | x_s = e_j)$, where $e_i, e_j \in \{0, 1\}^K$ are one-hot encoded vectors.

In the forward diffusion process, noise is added to observed data until it is indistinguishable from a stationary noise distribution. Zhao et al. (2024) show that the forward diffusion process for discrete-time and continuous-time diffusion can be unified into a single forward diffusion process defined by the accumulated transition matrix $\bar{Q}_{t|s}$, with $t > s$:

$$\bar{Q}_{t|s} = \bar{\alpha}_{t|s}I + (1 - \bar{\alpha}_{t|s})1m^\top, \quad (1)$$

where m is a known probability distribution vector such that $x_T \sim \mathcal{C}(x_T; m)$, 1 is vector of 1s, and $\bar{\alpha}_{t|s}$ is the accumulated noise schedule between time t and time s .

In the backward diffusion process, diffusion models aim to generate samples from $p_\theta(x_s|x_t)$ for some $0 < s < t \leq T$. This is commonly accomplished by taking $p_\theta(x_s|x_t) \approx q(x_s|x_t, x_0)$, where $q(x_s|x_t, x_0)$ is the s -step posterior distribution (Hoogeboom et al., 2021; Austin et al., 2021). Because x_0 is unknown, it is approximated by training a neural network model f_t^θ such that $p_\theta(x_0|x_t) = \mathcal{C}(x_0; f_t^\theta(x_t))$. Thus, $p_\theta(x_s|x_t)$ can be estimated by sampling $x_s \sim q(x_s|x_t, f_t^\theta(x_t))$. Analytically, this can be calculated as

$$q(x_s|x_t, f_t^\theta(x_t)) = \mathcal{C}\left(x_s; \frac{\bar{Q}_{t|s}x_t \odot \bar{Q}_s^\top f_t^\theta(x_t)}{x_t^\top \bar{Q}_t^\top f_t^\theta(x_t)}\right). \quad (2)$$

This posterior calculation can be further reduced to simplify computation (Zhao et al., 2024), which we follow in our work as well. We refer readers to Zhao et al. (2024) for a more detailed overview of the unified diffusion methodology.

The primary point of difference between discrete-time and continuous-time diffusion models now lies in the loss functions. We train both continuous-time and discrete-time diffusion models using the simplified variational lower bound loss functions provided by Zhao et al. (2024) alongside a cross-entropy loss.

3.2 DIFFER ADAPTED UNIFIED DIFFUSION

DIFFER (Current et al., 2024) incorporates the target product as an additional constraint to the diffusion generation process, akin to text-to-image generative diffusion models. Let $y_0^{1:L_y} \in \{0, 1\}^{L_y \times K}$ be the target product sequence associated with the reactant sequence $x_0^{1:L_x}$. Here, both y_0 and x_0 are SMILES strings representing molecular structures, with x_0 containing one or more concatenated reactant SMILES. Because the target product is known in retrosynthesis prediction, The diffusion problem can be parameterized such that

$$p_\theta(x_0|x_t, y_0) = \mathcal{C}(x_0; f_t^\theta(x_t, y_0)). \quad (3)$$

This formulation guides the diffusion process by incorporating the target product as an additional input to the model. Here, f_t^θ usually takes the form of an encoder-decoder model, where the encoder processes y_0 and the decoder acts as the diffusion component on x_t .

Notably, one issue remains in that the length of the reactant sequence L_x is unknown a priori, which is necessary to properly instantiate $x_T \sim \mathcal{C}(x_T; m)$. Inspired by MaskPredict (Ghazvininejad et al., 2019), DIFFER (Current et al., 2024) incorporates a length prediction component into the encoder half of f_t^θ , such that L_x can be estimated using some model $f_\ell^\phi : \{0, 1\}^{L \times K} \rightarrow \mathbb{R}^L$, so that $L_x \approx \arg \max f_\ell^\phi(y_0)$, where, L is an arbitrary maximum sequence length such that $L_x < L$ and $L_y < L$ for all x_0, y_0 in the data distribution. For the chemical retrosynthesis task, L_x and L_y are often highly correlated: thus, to encourage active learning, f_ℓ^ϕ is instead trained to predict the difference between L_x and L_y , i.e.,

$$L_x = L_y + (\arg \max f_\ell^\phi(y_0) - \Gamma), \quad (4)$$

where $\Gamma < L$ is a sufficiently large constant allowing f_ℓ^ϕ to effectively predict negative length changes. In practice, we set $\Gamma = \frac{L}{2}$, noting that the change in length will never be larger than half of the product sequence.

Current et al. (2024) show experimentally that if f_ℓ^ϕ is a perfect length-prediction model, then discrete-time diffusion outperforms all other methodologies for top-1 predictions by a significant margin, exhibiting that sequence length information can be highly informative when predicting reactant SMILES. However, the authors note that in practice, f_ℓ^ϕ cannot be sufficiently trained to accomplish such a notable performance increase due to 1.) the distribution of $(L_x - L_y)$ is heavily biased, 2.) the SMILES encoding of a molecule can produce sequences of varying length, forcing $f_\ell^\phi(y_0)$ to learn a one-to-many mapping, and 3.) accurate length prediction requires complex understanding of chemical reagents, molecular reactivity, and molecular structure, and is itself a nontrivial problem.

To remedy problems 1. and 2., DIFFER (Current et al., 2024) randomly appends $L_p \sim \mathcal{U}\{1, \kappa\}$ padding tokens to the SMILES representations of reactants, where κ is a set hyperparameter. This

padding procedure smooths the distribution of $(L_x - L_y)$ and allows flexibility in the output size of generated reactants. A detailed overview of this procedure is provided in Appendix A.1. To reduce reliance on the choice of κ , Differ trains multiple diffusion models $f_t^{\theta_\kappa}$ with different κ , producing an ensemble of models which generate possible reactants. This approach provides better coverage of the posterior distribution than a single model. We follow this same approach, but additionally incorporate continuous-time models and classifier-based guidance sampling into the ensemble.

3.3 SMILES ADAPTED PARTICLE GUIDANCE FOR CHEMICAL GENERATION

Aside from the incorporation of continuous-time diffusion models into the Differ methodology and the formalization of said methodology, our other primary motivation of this work is to explore the usage of classifier-based guidance mechanisms on the Differ ensemble models to improve output diversity. Classifier-based guidance steers diffusion sampling toward a specific classification c by sampling from a tempered distribution $p^\gamma(x_s|c, x_t) \propto p(c|x_s)^\gamma p_\theta(x_s|x_t)$, where $p_\theta(x_s|x_t)$ is the diffusion model, $p(c|x_s)$ is a classifier, and γ is the temperature parameter (Schiff et al., 2024):

$$\log p^\gamma(x_s|c, x_t) = \gamma \log p(c|x_s, x_t) + \log p(x_s|x_t). \quad (5)$$

Schiff et al. (2024) formalize classifier-based guidance for discrete-space diffusion specifically, with the assumption that $p^\gamma(x_s|c, x_t)$ factorizes independently across tokens. We follow their work to apply classifier-based guidance to our models.

3.3.1 SMILES ADAPTED PARTICLE GUIDANCE

One of the primary limitations of employing diffusion models for highly conditioned molecular generation tasks lies in the over-sampling of posterior peaks (Current et al., 2024; Sun et al., 2025), resulting in reduced diversity and number of unique generated molecules. To this end, we take inspiration from Corso et al. (2023) to encourage greater output diversity during sampling through guidance. Corso et al. apply the gradient of a potential function $\log \Phi_t(x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(n)})$ to encourage greater output diversity, where $x_t^{(i)}$ are independent samples from the diffusion model at time t . The potential function is constructed such that higher scores are given to distant point pairs:

$$\log \Phi_t(x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(n)}) = \frac{\bar{\alpha}_t}{2} \sum_{i,j} \mathcal{K}(x_t^{(i)}, x_t^{(j)}), \quad (6)$$

where $\mathcal{K}$ is a pairwise similarity kernel. Furthermore, the potential function is weighted by the noise schedule $\bar{\alpha}_t$ such that the guidance effect is stronger at earlier timesteps rather than later ones. Unfortunately, it is non-trivial to construct a differentiable similarity kernel between noised SMILES strings due to the non-singular nature of SMILES representations: the same molecule may be represented by many different SMILES sequences of various atom orderings.

To remedy this, we construct the potential function such that it maximizes the probability that $X_t = \{x_t^{(1)}, x_t^{(2)}, \dots, x_t^{(n)}\}$ are noisy variants of different molecules. Let $m_0^{(i)}$ be the molecular form of the denoised sequence $x_0^{(i)}$. Then we aim to maximize the probability

$$\bigcup_{i \neq j} p(m_0^{(i)} \neq m_0^{(j)} | x_t^{(i)}, x_t^{(j)}) = \prod_{i \neq j} p(m_0^{(i)} \neq m_0^{(j)} | x_t^{(i)}, x_t^{(j)}), \quad (7)$$

Which factors as a result of the independence of generated SMILES. Taking the log, we can construct the potential function as

$$\log \Phi_t(X_t) = \sum_{i \neq j} \log p(m_0^{(i)} \neq m_0^{(j)} | x_t^{(i)}, x_t^{(j)}). \quad (8)$$

We train a classifier to estimate $p(m_0^{(i)} \neq m_0^{(j)} | x_t^{(i)}, x_t^{(j)})$. Recognizing that the parameterized diffusion model f_t^θ is already trained to produce an estimate of the de-noised sequences, we train an additional classifier f_K^Φ to detect if $\hat{x}_{0,t}^{(i)} = f_t^\theta(x_t^{(i)})$ and $\hat{x}_{0,t}^{(j)} = f_t^\theta(x_t^{(j)})$ represent the same denoised molecular structure, i.e.,

$$p(m_0^{(i)} \neq m_0^{(j)} | x_t^{(i)}, x_t^{(j)}) = f_K^\Phi(\hat{x}_{0,t}^{(i)}, \hat{x}_{0,t}^{(j)}) = (f_K^\Phi \circ f_t^\theta)(x_t^{(i)}, x_t^{(j)}). \quad (9)$$

This parameterization results in the final potential function

$$\log \Phi_t(X_t) = \sum_{i \neq j} \log \left((f_K^\Phi \circ f_t^\theta)(x_t^{(i)}, x_t^{(j)}) \right). \quad (10)$$

3.3.2 APPLYING POTENTIAL FUNCTIONS VIA CLASSIFIER-BASED GUIDANCE

Rather than using this potential function directly, we instead construct a classifier on $x_t^{(i)}$ alone, recognizing that applying the classifier individually for all $x_t^{(i)} \in X_t$ reproduces the potential function:

$$\log p_{\Phi}(\mathbb{I}[\hat{m}_{0,t}^{(i)} \notin \hat{M}_t / \{\hat{m}_{0,t}^{(i)}\}] | x_t^{(i)}, X_t) = \sum_{j \neq i} \log \left((f_{\mathcal{K}}^{\Phi} \circ f_t^{\theta})(x_t^{(i)}, x_t^{(j)}) \right). \quad (11)$$

Here, $\hat{m}_{0,t}^{(i)}$ is the molecular form of $\hat{x}_{0,t}^{(i)}$, $\hat{M}_t$ is the set of molecules $\{\hat{m}_{0,t}^{(1)}, \dots, \hat{m}_{0,t}^{(n)}\}$ associated with X_t , and $\mathbb{I}$ is the indicator function. For the sake of clarity, we will refer to $\mathbb{I}[\hat{m}_{0,t}^{(i)} \notin \hat{M}_t / \{\hat{m}_{0,t}^{(i)}\}]$ as the guidance class c . We add the further stipulation that if $\hat{x}_{0,t}^{(i)}$ is not a valid SMILES string or if $\hat{m}_{0,t}^{(i)}$ is not a valid molecule, then the indicator function defaults to False.

We finalize our guidance strategy by follow the approach of Schiff et al. (2024) to apply classifier guidance to discrete sequences. This results in the posterior estimate

$$\log p_{\Phi, \theta}^{\gamma}(x_s | c, x_t) = \sum_{\ell=1}^L \log \left(\frac{p_{\Phi}(c | \tilde{x})^{\gamma} p_{\theta}(x_s^{[\ell]} | x_t)}{\sum_{\tilde{x}} p_{\Phi}(c | \tilde{x})^{\gamma} p_{\theta}(x_s^{[\ell]} | x_t)} \right), \quad (12)$$

where $\tilde{x} = [x_t^{1:\ell-1}, x_s^{\ell}, x_t^{\ell+1:L}]$ is the sequence such that $\tilde{x}$ and x_t are equal at all positions $\ell' \neq \ell$ and $\tilde{x}$ and x_s are equal at position ℓ . We set $\gamma = \bar{\alpha}_t$ to weight the guidance effect similarly to Equation 6. Finally, we use the first-order Taylor approximation of $\log p_{\Phi}$ to efficiently compute the classifier $p_{\Phi}(c | \tilde{x})$ using only a single forward pass (Schiff et al., 2024):

$$\log p_{\Phi}(c | \tilde{x}) = (\tilde{x} - x_t)^{\top} \nabla_{x_t} \log p_{\Phi}(c | x_t) + \log p_{\Phi}(c | x_t). \quad (13)$$

We refer readers to Schiff et al. (2024) for detailed proofs of this result.

Training and implementation details for $f_{\mathcal{K}}^{\Phi}$ are provided in Appendix A.2. During inference, $f_{\mathcal{K}}^{\Phi}$ is incorporated in a manner reminiscent of reinforcement learning, allowing the model to learn and adapt throughout the diffusion process.

4 EXPERIMENTS

Experimental Setup. We train and test our models on the USPTO-50K dataset (Lowe, 2017), a commonly used set of patented reactions that provides SMILES strings for both reactants and products and widely regarded as the benchmark dataset for retrosynthesis prediction. Reaction solvents are not included in the retrosynthesis task. Each pair of products and reactants is augmented 20 times according to Zhong et al. (2022), producing a final dataset of 1M root-aligned SMILES pairs. We use the augmented train, validation, and test sets provided by Zhong et al.¹

The DIFFER² ensemble contains eight individually trained diffusion models, depicted in Figure 1. Each model shares the same network architecture, differentiated only by the diffusion parameters and padding augmentation described in Section 3.2. Four of the diffusion models follow the original DIFFER training (Current et al., 2024) with losses adjusted according to Zhao et al. (2024): these models operate in discrete-time with padding parameter $\kappa \in \{20, 40, 60, 80\}$. The remaining four models operate in continuous-time, each paralleling their discrete-time counterparts with a padding parameter $\kappa \in \{20, 40, 60, 80\}$. Further details on model architecture are provided in Appendix A.3. Models which include particle guidance described in Section 3.3 are denoted DIFFER²PG and DIFFER²PG+. In DIFFER²PG, guidance sampling is run instead of the standard sampling procedure. In DIFFER²PG+, guidance sampling is run alongside standard sampling procedures, resulting in an output distribution that combines DIFFER² and DIFFER²PG.

Evaluation Metrics. We evaluate our models using standard top- k accuracy metrics. The DIFFER² ensemble is run on each product in the test set and generated reactants are ranked by frequency, with the most generated reactant having the highest ranking. The top- k accuracy is defined as the proportion of reactions in which the ground-truth reactant is in the top- k ranks. Top- k accuracy is

¹<https://github.com/otori-bird/retrosynthesis>

reported for $k \in \{1, 3, 5, 10\}$. We compare the DiffER² ensemble with a wide range of template-based, semi-template, and template-free methods for retrosynthesis prediction. However, we primarily focus on comparing performance against other template-free methods. In addition to top-k accuracy, we include results for round-trip accuracy using a forward synthesis model in Appendix ?? and provide a case study on a ring-forming reaction in the main results and three other reactions in Appendix A.8.

Property-Specific Performance Analysis. We additionally analyze how model performance compares against various structural properties of the SMILES sequences to better understand model performance. These properties include the target length difference between product and reactant SMILES, the edit distance between product and reactant SMILES, if the target reaction is ring adding/ring removing, and the number of branches in the target SMILES, each detailed in Appendix A.7. For each property, we calculate the slope and significance of a regression between the property statistic and the top-k accuracy. In addition to these results, we report the performance of DiffER²PG across different reaction types in Appendix A.6. We analyze how different molecular structures correlate with model performance in Appendix A.9. Finally, we analyze the attention maps for a non-ensemble diffusion model in Appendix A.10.

5 RESULTS

Top-K accuracies for the DiffER² ensemble are reported in Table 1. DiffER²PG+ achieves the best top-1, top-3, and top-5 (statistically tied with R-SMILES) among template-free models and the second-best top-10 accuracy behind R-SMILES (Zhong et al., 2022). R-SMILES achieves the best top-5 and top-10 performance, but lags behind DiffER²PG+ for $k = 1, 3$. DiffER²PG+ may better estimate posterior peaks for SMILES sequences than R-SMILES, but it still struggles to sample from lower probability regions in the posterior. This result is paralleled in GDiffRetro (Sun et al., 2025), a semi-template method that utilizes diffusion sampling techniques. GDiffRetro achieves remarkable $k = 1$ accuracy, the highest of all baseline models, but struggles to produce strong results for higher values of k compared to other methods. Sun et al. (2025) suggest this may be due to the oversampling of posterior peaks, which is also observed in our models. By employing an ensemble of diffusion models and incorporating SMILES-adapted particle guidance, DiffER²PG+ can mitigate some of this bias and provide greater coverage of the posterior, as evidenced by our stronger performance for $k > 1$ compared to GDiffRetro. Compared to the original DiffER (Current et al., 2024), the DiffER² ensembles all achieve stronger results than the discrete-time only ensemble, particularly for higher values of k . This suggests that including continuous-time models greatly improves the diversity of output reactants. Notably, there is minimal difference between all DiffER methodologies for top-1 accuracy, indicating that the location of posterior peaks for the various ensembles may be similar, but that the shape of the distribution for lower probability samples is different. Interestingly, the top-k performances of DiffER² and DiffER²PG are remarkably similar despite the latter ensemble producing more reactants.

Figure 2 shows box plots of the unique reactants generated by each ensemble for products in the test set. The greater number of unique molecules produced by DiffER²PG indicates that the

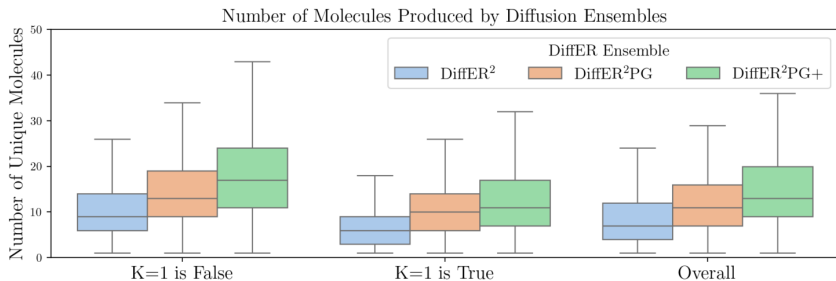


Figure 2: Box plot showing the number of unique molecules produced by the various DiffER² ensembles. "K=1 is False" indicates reactions that do not have an accurate top-1 reactant. "K=1 is True" indicates reactions that do have an accurate top-1 reactant. "Overall" includes all reactions.

Table 1: Top-K accuracy for a select set of baseline models. A full table of results can be found in Appendix A.4. The best performing model in each category is **bolded**, while the second-best performing model is underlined. A † indicates results that are statistically significant from DiffER at the 95% level.

Category	Model	K=1	3	5	10
Template-based	GLN (Dai et al., 2019)	<u>52.5</u>	<u>69.0</u>	<u>75.6</u>	<u>83.7</u>
	LocalRetro (Chen & Jung, 2021)	53.4	77.5	85.9	92.4
Semi-template	RetroPrime (Wang et al., 2021)	51.4	70.8	74.0	<u>76.1</u>
	G ² Retro (Chen et al., 2023)	<u>53.9</u>	<u>74.6</u>	<u>80.7</u>	86.6
	GDiffRetro (Sun et al., 2025)	58.9	79.1	81.9	-
Template-free	MEGAN (Sacha et al., 2021)	48.1	70.7	78.4	86.1
	R-SMILES (Zhong et al., 2022)	56.3	79.2	86.2	91.0
	DiffER (Current et al., 2024)	<u>57.6</u>	79.0	84.1	87.4
	DiffER ²	<u>57.6</u>	<u>79.5</u>	85.3 [†]	88.2
	DiffER ² PG	<u>57.4</u>	79.4	85.4 [†]	88.6 [†]
	DiffER ² PG+	57.7	79.8	<u>86.1</u> [†]	<u>89.9</u> [†]

SMILES-adapted particle guidance methodology does result in greater output diversity and reduced sampling of posterior peaks. The combined result of DiffER²PG+ has an even greater number of unique molecules, indicating that the difference between sample sets generated from DiffER² and DiffER²PG is substantial. Notably, this result may be impacted by the limited number of samples taken from each model in the ensemble, which may not reflect the full posterior distribution for each model. Regardless, DiffER²PG is an effective approach to improving output diversity while maintaining comparable performance to DiffER². This is desirable because multiple reactants could produce the same product, thus validating accuracy, and output diversity is critical. Like Current et al. (2024), we observe that fewer molecules are sampled when the top-1 result is accurate, indicating that the models still over-sample from posterior peaks to the deficit of output diversity. However, these effects are lessened compared to DiffER.

Case Study. A case study on an N-heterocyclization reaction with dihalides for the DiffER²PG+ ensemble is presented in Figure 3. Additional case studies are presented in Appendix A.8. The target reaction is a reaction of a primary amine with a dihalide in an alkaline aqueous medium to form a heterocyclic ring (Ju & Varma, 2006). While the ensemble model predicts this reaction, it is only predicted 4% of the time, making it the third most frequently predicted reaction. The top-ranked predicted reaction is a Borch reductive amination (Borch, 2003) in which a carbonyl group is reacted with a secondary amine in the presence of a reducing agent to produce a tertiary amine. Unfortunately, due to the presence of multiple carbonyl groups in the predicted reactant, this reaction will likely produce unwanted byproducts. The second-ranked reaction is a substitution reaction with a mesylate ester and a secondary amide nucleophile. Again, this reaction does produce the desired product, but may produce byproducts due to the electrophilicity of the carbamate group, which may react with the amine nucleophile.

To recap, all three top-ranked reactions could be used to form the target product, but only the 3rd reaction, which matches the ground truth, would limit the creation of byproducts. Notably, this reactant is not produced by the original DiffER ensemble (Current et al., 2024), demonstrating that DiffER²PG+ significantly improves output diversity to benefit performance. In this case study, as well as those presented in Appendix A.8, DiffER²PG+ demonstrates significant capability to predict reactants that could be used to produce the desired product, though some reactions may be suboptimal due to the possibility of undesirable side products. Thus, while the models can learn appropriate reaction mechanisms and patterns, the predicted forward reaction’s purity may not be considered by the models. Future work should consider approaches that include reducing the possibility of side products as an objective. Furthermore, ambiguity in the optimality of reactants could be mitigated by providing additional guidance to the model in a human-in-the-loop setting.

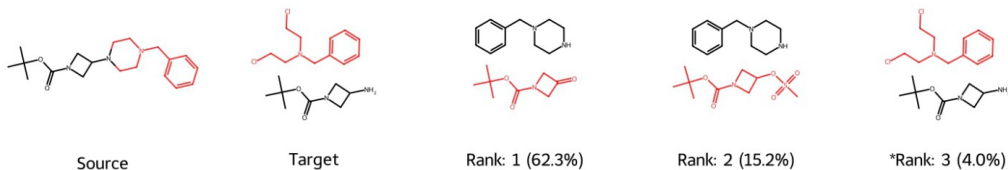


Figure 3: Case study on a N-heterocyclization reaction with dihalides. From left to right: the source product, the ground-truth reactant(s), and the top-3 predicted reactant(s) for the Differ²PG+ ensemble. Red highlighting indicates differences between the source product and the reactants. The value in parentheses indicates the rate at which the reactant(s) are produced by the ensemble.

Property Results. Figure 4 shows the results of the property-specific performance analysis described in Section 4 for top-1 accuracy, with additional results available in Appendix A.7. These results highlight substantial differences in the capabilities of Differ²PG+ and R-SMILES. Differ²PG+ is significantly more capable at predicting reactant SMILES with a larger increase in sequence length than the auto-regressive model. This result also helps explain the increased capability of diffusion models to predict ring-removing reactions, as these are often reactions with a high increase in sequence length. This difference in capability can be understood via the attention maps for the diffusion models: rather than building up the sequence token-by-token like an auto-regressive model, the diffusion models work backward, attending to the full-sequence length in early layers then decreasing the attended length as the model goes deeper (see Appendix A.10). Thus, diffusion models may be "less inclined" to produce shorter sequences.

In contrast, diffusion models seem less capable of modeling ring-adding reactions than auto-regressive models, though this difference is only significant for $k > 1$. Both models have a negative association with the edit distance between product and reactant SMILES, except for top-1 accuracy, in which Differ²PG+ has no significant relationship with the edit distance. This suggests that some reactions with large edits remain "easy" for Differ²PG+ to predict. Both models struggle to generate sequences with highly non-sequential behavior, as evidenced by the negative relationship between performance and branch count. Overall, Differ²PG+ outperforms R-SMILES conditioned on all properties except for ring-adding reactions.

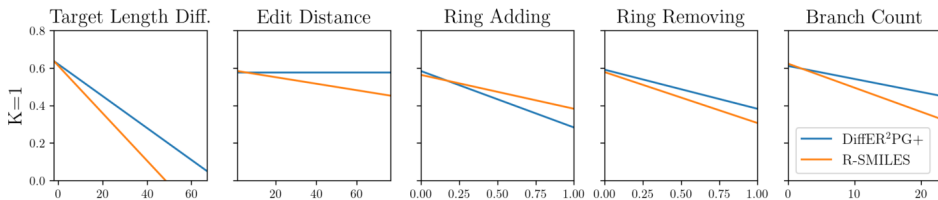


Figure 4: Lines of best fit between top-k accuracy and reactant properties for the Differ²PG+ and R-SMILES (Zhong et al., 2022) models. All slopes are significantly different from 0 ($p < 0.01$) except Edit Distance.

6 CONCLUSION

In this work, we introduce SMILES-adapted particle guidance to encourage output diversity in diffusion models for SMILES generation on the chemical retrosynthesis task. Combining this guidance mechanism with the ensemble approach of Differ (Current et al., 2024), we can significantly increase the number of unique molecules generated while improving top-k accuracy metrics on benchmark datasets. Furthermore, we offer an in-depth empirical analysis on the capabilities and limitations of discrete diffusion for SMILES generation compared to traditional auto-regressive approaches. In future work, we aim to investigate additional guidance methods for SMILES diffusion models to reduce byproducts and explore applications of rectified linear flows for SMILES generation.

REPRODUCIBILITY

We detail all model architectures and hyperparameters used to produce DIFFER² in Appendix A.3. We follow the methods outlined in Section 3 and Appendix A.2 alongside the experimental setup in Section 4 to train and evaluate our models. All code used to produce the results of DIFFER² will be made publicly available.

REFERENCES

- Amira Alakhdar, Barnabas Poczos, and Newell Washburn. Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, 64(19):7238–7256, 2024.
- 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.
- Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.
- Richard F Borch. Reductive amination with sodium cyanoborohydride: N, n-dimethylcyclohexylamine: Cyclohexanamine, 4, 4-dimethyl-. *Organic Syntheses*, 52: 124–124, 2003.
- Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.
- Shuan Chen and Yousung Jung. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au*, 1(10):1612–1620, 2021.
- Ziqi Chen, Oluwatosin R Ayinde, James R Fuchs, Huan Sun, and Xia Ning. G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Communications Chemistry*, 6(1):102, 2023.
- Clifford C. Clogg, Eva Petkova, and Adamantios Haritou. Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, 100(5):1261–1293, 1995. ISSN 00029602, 15375390. URL <http://www.jstor.org/stable/2782277>.
- 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.
- Gabriele Corso, Yilun Xu, Valentin De Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-iid diverse sampling with diffusion models. *arXiv preprint arXiv:2310.13102*, 2023.
- Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869, 2023.
- Sean Current, Ziqi Chen, Xia Ning, et al. Differ: Categorical diffusion models for chemical retrosynthesis. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024.
- Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems*, 32, 2019.
- Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.
- Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. *arXiv preprint arXiv:1904.09324*, 2019.

- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. *arXiv preprint arXiv:2310.05793*, 2023.
- Nate Gruver, Samuel Stanton, Nathan Frey, Tim GJ Rudner, Isidro Hotzel, Julien Lafrance-Vanasse, Arvind Rajpal, Kyunghyun Cho, and Andrew G Wilson. Protein design with guided discrete diffusion. *Advances in neural information processing systems*, 36:12489–12517, 2023.
- Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022.
- Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusion-bert: Improving generative masked language models with diffusion models. *arXiv preprint arXiv:2211.15029*, 2022.
- Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.
- Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Emiel Hoogetboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021.
- Emiel Hoogetboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.
- 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.
- Yuhong Ju and Rajender S Varma. Aqueous n-heterocyclization of primary amines and hydrazines with dihalides: microwave-assisted syntheses of n-azacycloalkanes, isoindole, pyrazole, pyrazolidine, and phthalazine derivatives. *The Journal of organic chemistry*, 71(1):135–141, 2006.
- 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.
- Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alán Aspuru-Guzik. SELFIES: a robust representation of semantically constrained graphs with an example application in chemistry. *CoRR*, abs/1905.13741, 2019. URL <http://arxiv.org/abs/1905.13741>.
- Miguelangel Leon, Yuriy Perezhohin, Fernando Peres, Aleš Popovič, and Mauro Castelli. Comparing smiles and selfies tokenization for enhanced chemical language modeling. *Scientific Reports*, 14(1):25016, 2024.
- Bowen Liu, Bharath Ramsundar, Prasad Kawthekar, Jade Shi, Joseph Gomes, Quang Luu Nguyen, Stephen Ho, Jack Sloane, Paul Wender, and Vijay Pande. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science*, 3(10):1103–1113, 2017.
- Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 289–299, 2023.

- Daniel Lowe. Chemical reactions from US patents (1976-Sep2016). 6 2017. doi: 10.6084/m9.figshare.5104873.v1. URL https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.
- Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.
- Ali Alipour Najmi, Rainer Bischoff, and Hjalmar P Permentier. N-dealkylation of amines. *Molecules*, 27(10):3293, 2022.
- Bo Ni, David L Kaplan, and Markus J Buehler. Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem*, 9(7):1828–1849, 2023.
- C Ramesh, N Ravindranath, and Biswanath Das. Simple, efficient, and selective deprotection of phenolic methoxymethyl ethers using silica-supported sodium hydrogen sulfate as a heterogeneous catalyst. *The Journal of Organic Chemistry*, 68(18):7101–7103, 2003.
- 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.
- Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. *arXiv preprint arXiv:2412.10193*, 2024.
- 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.
- Arne Schneuing, Charles Harris, Yuanqi Du, Kieran Didi, Arian Jamasb, Ilia Igashov, Weitao Du, Carla Gomes, Tom L Blundell, Pietro Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.
- Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry—A European Journal*, 23(25):5966–5971, 2017.
- 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.
- Yuning Shen, Julia E Borowski, Melissa A Hardy, Richmond Sarpong, Abigail G Doyle, and Tim Cernak. Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers*, 1(1):1–23, 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 Coley, Andreas Krause, and Regina Barzilay. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems*, 34:9405–9415, 2021.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Dean Sumner, Jiazhen He, Amol Thakkar, Ola Engkvist, and Esben Jannik Bjerrum. Levenshtein augmentation improves performance of smiles based deep-learning synthesis prediction. 2020.
- Ruoxi Sun, Hanjun Dai, Li Li, Steven Kearnes, and Bo Dai. Towards understanding retrosynthesis by energy-based models. *Advances in Neural Information Processing Systems*, 34:10186–10194, 2021.
- Shengyin Sun, Wenhao Yu, Yuxiang Ren, Weitao Du, Liwei Liu, Xuechang Zhang, Ying Hu, and Chen Ma. Gdiffretro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. *arXiv preprint arXiv:2501.08001*, 2025.
- Zhengcai 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.
- 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.
- Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *BioRxiv*, pp. 2022–12, 2022.
- David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.
- Minkai Xu, Alexander S Powers, Ron O Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*, pp. 38592–38610. PMLR, 2023.
- Zhe Xu, Ruizhong Qiu, Yuzhong Chen, Huiyuan Chen, Xiran Fan, Menghai Pan, Zhichen Zeng, Mahashweta Das, and Hanghang Tong. Discrete-state continuous-time diffusion for graph generation. *arXiv preprint arXiv:2405.11416*, 2024.
- 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.
- Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.
- Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Fei Huang, and Songfang Huang. Seqdiffuseq: Text diffusion with encoder-decoder transformers. *arXiv preprint arXiv:2212.10325*, 2022.

- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.
- Lingxiao Zhao, Xueying Ding, Lijun Yu, and Leman Akoglu. Unified discrete diffusion for categorical data. *arXiv preprint arXiv:2402.03701*, 2024.
- 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.