

RESEARCH

Open Access



DIFFER: categorical diffusion ensembles for single-step chemical retrosynthesis

Sean Current^{1*}, Ziqi Chen¹, Daniel Adu-Ampratwum², Xia Ning^{1,2,3,4} and Srinivasan Parthasarathy^{1*}

Abstract

Methods for automatic chemical retrosynthesis have found recent success through the application of models traditionally built for natural language processing, primarily through transformer neural networks. These models have demonstrated significant ability to translate between the SMILES encodings of chemical products and reactants, but are constrained as a result of their autoregressive nature. We propose DIFFER, an alternative template-free method for single-step retrosynthesis prediction in the form of categorical diffusion, which allows the entire output SMILES sequence to be predicted in unison. We construct an ensemble of diffusion models which achieves state-of-the-art performance for top-1 accuracy and competitive performance for top-3, top-5, and top-10 accuracy among template-free methods. We prove that DIFFER is a strong baseline for a new class of template-free model and is capable of learning a variety of synthetic techniques used in laboratory settings.

Scientific contribution

DIFFER represents a novel template-free technique for single-step retrosynthesis prediction which is able to outperform a variety of other template-free methods on top-k accuracy metrics. By constructing an ensemble of categorical diffusion models with a novel length prediction component with variance, our method is able to approximately sample from the posterior distribution of reactants, producing results with strong metrics of confidence and likelihood. Furthermore, our analyses demonstrate that accurate prediction of the SMILES sequence length is key to further boosting the performance of categorical diffusion models.

Retrosynthesis prediction is an important step in organic synthesis tasks, particularly those posed for drug discovery or drug engineering. In forward synthesis prediction, the products of a chemical reaction are predicted from known reactants; retrosynthesis prediction reverses this process, instead predicting possible reactants that would produce a target product. Repeated application of retrosynthesis prediction can help chemists construct synthetic pathways for drug targets [1–3], promoting the discovery and advancement of new pharmaceuticals. While numerous types of models for computer-aided retrosynthesis have been proposed, many modern approaches utilize data-driven machine learning to find suitable models for retrosynthesis prediction.

*Correspondence:

Sean Current
current.33@osu.edu
Srinivasan Parthasarathy
srini@cse.ohio-state.edu

¹ Computer Science and Engineering, The Ohio State University, Columbus 43210, OH, USA

² College of Pharmacy, The Ohio State University, Columbus 43210, OH, USA

³ Translational Data Analytics Institute, The Ohio State University, Columbus 43210, OH, USA

⁴ Biomedical Informatics, The Ohio State University, Columbus 43210, OH, USA



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Template-free retrosynthesis models are one such data-driven approach, which aim to directly learn transitions between product and reactant molecules. In template-free approaches, entire reactants are generated by the model, rather than explicitly editing existing structures in the product molecule to form reactants. Recent advances in template-free machine learning models for chemical retrosynthesis have taken advantage of transformer architectures originally crafted for natural language tasks. Instead of operating on natural language, many of these models are set to operate on SMILES [4] encodings of chemical products and reactants, effectively translating between two molecular sets [5]. Indeed, recent work has done precisely this, going as far as using neural machine translation models originally built for language tasks for retrosynthesis prediction [6, 7]. A more detailed overview of current work in single-step retrosynthesis is provided in Appendix 1. While these models exhibit remarkable performance compared to other baseline methods for

retrosynthesis, they impose autoregressive constraints during training and prediction, enforcing sequential decoding of the SMILES string.

In this work, we offer *DiffER*, Diffusion Ensembles for Single-Step Chemical Retrosynthesis. *DiffER* offers an alternative to sequential decoding of SMILES strings in the prediction process by utilizing categorical diffusion models rather than autoregressive models. This exchange allows *DiffER* to decode the entire SMILES string in unison, rather than in an autoregressive manner. We hypothesize that by allowing the model to predict the entire SMILES string in unison, it may better learn structural relationships within the molecule. Additionally, diffusion models have experienced a recent surge of success for many generation tasks across domains. By employing categorical diffusion for the chemical retrosynthesis task, we provide a strong and competitive baseline for a new class of model. An outline of our approach is displayed in Fig. 1, while a more detailed overview is available in Sect. Methods. *DiffER* achieves state-of-the-art top-1

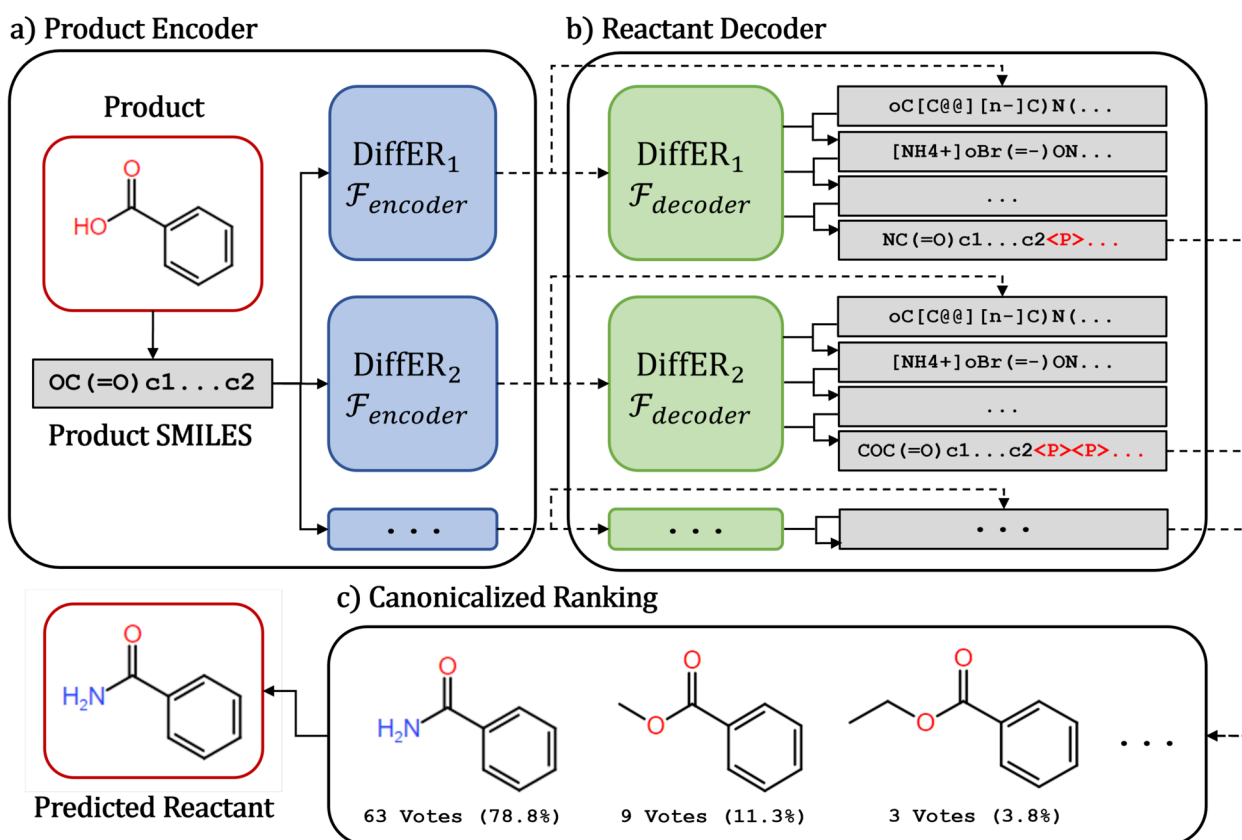


Fig. 1 Schematic of *DiffER*, our diffusion process for retrosynthesis. **a** The Product Encoder, which converts the product to a random, root-aligned SMILES form and encodes it using the encoder portion of the *DiffER* models. The encoder models also predict the size of the diffusion noise vector. **b** The Reactant Decoder, which diffuses the reactant from randomly initialized categorical noise using the decoder portion of the *DiffER* models. Decoder models can predict pad tokens $\langle P \rangle$ to diffuse SMILES strings of varying lengths. **c** The Canonicalized Ranking, which ranks the canonical molecules predicted by the *DiffER* models by the number of times they are generated. The top-ranked reactant is the predicted reactant.

accuracy and competitive performance for top-3, top-5, and top-10 accuracy among template-free methods.

Our main contributions to the domain of computer-aided chemical retrosynthesis are as follows:

- We develop and test **DiffER**, an ensemble of diffusion models for the chemical retrosynthesis task and compare our performance against a variety of existing methods, achieving state-of-the-art top-1 accuracy and competitive top-3 and top-5 accuracy.
- We implement a novel method for estimating the length of the output sequence prior to prediction so that the diffusion process can be properly initialized, while accounting for possible error and variance in the predicted sequence length.
- We compare against a diffusion model with oracle length prediction to demonstrate an experimental upper-limit of the diffusion approach if the output sequence length is able to be perfectly estimated. We find that the oracle model greatly outperforms all existing methods for computer-aided retrosynthesis, demonstrating the viability of diffusion models for this task given a target sequence length.
- We also include results for a baseline diffusion model with length prediction which does not account for variance or error in the output length, and show that our training and sampling methodology greatly improves the performance of diffusion models.
- We demonstrate the results of our method on a set of four reactions, highlighting the benefits of our approach as well as the limitations for each reaction.
- We discuss the current advantages and limitations of our diffusion approach and identify directions of future research to improve upon our work.

Methods

We define the retrosynthesis task as a sequence-to-sequence modeling problem and combine the approaches of Hooeboom et al. [8], DiffuSeq [9, 10], and Mask-Predict [11] to train conditional, multinomial diffusion models for the retrosynthesis task. We utilize a traditional encoder-decoder transformer architecture to construct individual conditional diffusion models, where the encoder is responsible for learning the representation of the conditional features and the decoder acts as the diffusion approximation model. In the end, eight individual diffusion models are ensembled to produce **DiffER**, a novel ensemble model for chemical retrosynthesis.

Both target reactants and source products are represented as SMILES [4] sequences, which encode molecular structures in a string format. Let $\mathcal{V}$ be the set of tokens in the SMILES vocabulary, representing atoms, bond types,

and other molecular structures. The reactant and product smiles can be represented as sequences $\mathbf{x}_0 \in \{0, 1\}^{L_x \times K}$ and $\mathbf{y}_0 \in \{0, 1\}^{L_y \times K}$, where $K = ||\mathcal{V}||$ is the number of tokens in the SMILES vocabulary and L_x , L_y represent the number of tokens used to construct the reactant and product SMILES sequences, respectively. We refer to the product SMILES as the “source” sequence, as it represents the conditional input to the diffusion model, and the reactant SMILES as the “target” sequence, as it is the generative goal of the diffusion model.

Let $\mathcal{C}(\cdot|k)$ denote a categorical probability distribution with parametrized probability k . The goal of the diffusion model is to transform a sequence of pure categorical noise $\mathbf{x}_T \in \{0, 1\}^{L_x \times K}$ into the target reactant SMILES $\mathbf{x}_0$ by iteratively denoising $\mathbf{x}_T$ over T timesteps, where $\mathbf{x}_T \sim \mathcal{C}(\mathbf{x}_T|\mathbf{m})$ is sampled from a known categorical distribution with probability vector $\mathbf{m}$. In this work, we assume the known distribution has uniform probability across all tokens, i.e., $\mathbf{m} = \mathbf{1}/K$, where $\mathbf{1}$ is a vector of ones. We let $\mathbf{x}_t \in \{0, 1\}^{L_x \times K}$ represent the noisy target reactant SMILES at diffusion step $t \in [0, T]$, and use the notation $\mathbf{x}_t^{[i]} \in \{0, 1\}^K$ to denote $\mathbf{x}_t$ indexed at the i th position. Finally, let $\mathcal{F}$ represent a transformer neural network with encoder $\mathcal{F}_{\text{encoder}}$, decoder $\mathcal{F}_{\text{decoder}}$, and sequence length prediction classifier $\mathcal{F}_{\text{length}}$.

Forward noising process

Following Hooeboom et al. [8], we define the multinomial diffusion process q as a categorical distribution $\mathcal{C}$ with probability β_t of sampling uniformly from $\{0, 1\}^K$ for any step in the sequence. We apply the forward diffusion process to sample the noisy $\mathbf{x}_t$ from $\mathbf{x}_{t-1}$:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{C}(\mathbf{x}_t|(1 - \beta_t)\mathbf{x}_{t-1} + \beta_t/K). \quad (1)$$

This forms a Markov chain allowing $\mathbf{x}_t$ to be directly sampled from clean target sequence $\mathbf{x}_0$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{C}(\mathbf{x}_t|\bar{\alpha}_t\mathbf{x}_0 + (1 - \bar{\alpha}_t)/K), \quad (2)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{\tau=1}^t \alpha_\tau$. Note that in the forward noising process, we only apply noise to the target reactant sequence $\mathbf{x}_0$ and not the source product sequence.

Conditional denoising

We continue following the work of Hooeboom et al. to denoise the categorical data. We construct the posterior distribution of the categorical model as

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{C}(\mathbf{x}_{t-1}|\theta_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0)), \quad (3)$$

where

$$\theta_{\text{post}}(\mathbf{x}_t, \mathbf{x}_0) = \bar{\theta} / \sum_{k=1}^K \hat{\theta}_k \quad (4)$$

and

$$\bar{\theta} = [\alpha_t \mathbf{x}_t + (1 - \alpha_t)/K] \odot [\hat{\alpha}_{t-1} \mathbf{x}_0 + (1 - \hat{\alpha}_{t-1})/K], \quad (5)$$

where $\odot$ is element-wise multiplication.

Notably, as $\mathbf{x}_0$ is not known during inference, it must be estimated using a neural network. Inspired by DiffuSeq [9, 10], we concatenate a representation of the source sequence to the noised representation of the target sequence to use as input to the $\mathbf{x}_0$ approximation model. In contrast to DiffuSeq, which uses the embedded representation of the source product sequence $\mathbf{y}_0$, we process the source sequence with a transformer encoder model prior to concatenating, which is used as the memory component in the decoder architecture [12]. Thus, the reverse diffusion process is fully modeled as

$$p_{\theta}(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{y}_0) = \mathcal{C}(\mathbf{x}_{t-1}|\theta_{\text{post}}(\mathbf{x}_t, \mathcal{F}_{\text{decoder}}(\mathbf{x}_t|\mathcal{F}_{\text{encoder}}(\mathbf{y}_0)))), \quad (6)$$

where $||$ is the concatenation operator. Additionally, we add a sinusoidal embedding [8] of the current diffusion timestep t to the diffusion model input $\mathbf{x}_t$, informing the model of the current diffusion timestep.

Length prediction

One downside of the diffusion methodology compared to normal autoregressive methods is the requirement that the total length of the sequence must be known prior to inference so that the starting uniform noise distribution can be properly initialized. Prior work, such as DiffuSeq [9, 10], have primarily approached this issue by padding the sequence with observable PAD tokens up to a maximum length, and allowing the diffusion model to predict the location of PAD tokens as necessary. In contrast, MaskPredict [11] presents the idea of a length prediction token, which is learned by the transformer encoder block and used to initialize the input to the transformer decoder. We similarly prepend a LENGTH token to the source sequence $\mathbf{y}_0$ and use the learned representation of the LENGTH token for the input product to predict a probability distribution vector $\ell_x \in [0, 1]^L$ over the possible sequence lengths from 1 up to some arbitrary maximum length L , such that $\arg \max \ell_x \approx L_x$ and $\sum_i \ell_x^{[i]} = 1$:

$$p(\ell_x|\mathbf{y}_0) = \mathcal{F}_{\text{length}}(\mathcal{F}_{\text{encoder}}(\mathbf{y}_0)). \quad (7)$$

Here, $\mathcal{F}_{\text{length}}$ is a feed-forward classifier function dependent only on the first column of $\mathcal{F}_{\text{encoder}}(\mathbf{y}_0)$, which represents the LENGTH token.

Upon initial testing, we find that the baseline implementation of MaskPredict length prediction performs poorly for this task due to its rigidity in predicting the sequence length and bias toward predicting shorter sequences than the ground truth. To remedy this, we propose a novel adjustment to the method by appending a uniform random number $n \sim \mathcal{U}(1, N)$ of observable padding tokens to the decoder sequences during training. This increases the expected length of the sequence by $\frac{(1+N)}{2}$ while allowing the diffusion model to predict pad tokens within a range of the target length. Notably, as the probability of predicting n padding tokens is equivalent for $n \in [1, N]$, the probability of predicting sequences with lengths within $\frac{(1+N)}{2}$ of the predicted length is approximately equal. This offers benefits of both the MaskPredict and DiffuSeq approaches: The diffusion model is provided with information about the target length from the encoder while still allowing for the possibility of predicting a variety of sequence lengths rather than only the predicted target length. Intuitively, lower limits of N place more focus on the length prediction component, while higher values of N allow the model to vary more from the initial length prediction. Additionally, as the lengths of the input product SMILES sequence and output reactant SMILES sequence of the retrosynthesis task are highly correlated, we shift the methodology to predict the difference in length between the source and target rather than the total length of the target, such that $\arg \max \ell_x \approx L_x - L_y + C$, where C is some arbitrary constant allowing for negative length differences.

Ensemble voting of models

In order to stabilize and diversify the output of the diffusion model, we implement an ensemble of various individual diffusion models with different values of the random padding limit N . During inference, multiple samples are drawn from each model trained on different N . Outputs are then ranked according to the number of times they are sampled overall. Ties are broken using a ranked-choice voting scheme. This process encourages output diversity by combining models trained with varying degrees of reliance on the length prediction component. We refer to the constructed ensemble model as `DiffE`.

We identically construct and train the individual models that compose `DiffE`, varying only the maximum amount of random padding that can be added on to the target sequence during training. We test various voting combinations of models with random padding limit $N \in \{5, 10, 15, 20, 30, 40, 50, 60, 70, 80, 90\}$. We find that models trained with $N < 20$ perform poorly, as they place too much focus on the length prediction component, which is biased toward a lower increase in length.

We construct the final ensemble using eight models trained on $N \in \{20, 30, 40, 50, 60, 70, 80, 90\}$.

Loss functions

We utilize a combination of mean square error (MSE) and variational lower bound (VLB) losses to train the diffusion models, the latter of which is derived using the Kullback–Leibler (KL) divergence according to Hoogetboom et al [8]. We apply MSE loss directly to the target reactant SMILES $\mathbf{x}_0$, while the VLB loss uses the sampled posterior of the predicted target:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_T \sim q} [\|\mathbf{x}_0^2 - \hat{\mathbf{x}}_0^2\|] \quad (8)$$

$$\mathcal{L}_{\text{VLB}} = \mathbb{E}_{\mathbf{x}_1, \dots, \mathbf{x}_T \sim q} \left[\sum_{k=0}^K \mathbf{x}_0^{[k]} \log \hat{\mathbf{x}}_0^{[k]} - \sum_{t=2}^T \text{KL}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{y}_0)) \right], \quad (9)$$

Where $\hat{\mathbf{x}}_0 = \mathcal{F}_{\text{decoder}}(\mathbf{x}_t | \mathcal{F}_{\text{encoder}}(\mathbf{y}_0))$ is the approximation of $\mathbf{x}_0$ obtained from the neural network $\mathcal{F}$. Finally, we include a loss term for the length prediction task using cross entropy:

$$\mathcal{L}_\ell = \mathbb{E}_{\ell_x \sim p} [\mathbf{e}_\ell^\top \log p(\ell_x | \mathbf{y}_0)], \quad (10)$$

where $\mathbf{e}_\ell \in \{0, 1\}^L$ is a one-hot encoded vector representing the length difference $L_x - L_y$ between the reactant sequence $\mathbf{x}_0$ and the product sequence $\mathbf{y}_0$. Additionally, when training the diffusion models, we employ the importance-based time sampling algorithm used in DiffuSeq [9, 10].

Reproducibility

We apply our models to the USPTO-50K [13] data set with Root-aligned SMILES augmentation [6], which greatly improves the performance of the diffusion models. We directly utilize the dataset splits provided by the authors of Root-aligned SMILES. We form `DIFFER` as an ensemble of models with random padding limit $N \in \{20, 30, 40, 50, 60, 70, 80, 90\}$. Each model takes the form of an encoder-decoder transformer architecture with 6 layers and 8 attention heads with hidden dimension size 512, feed-forward size 2048, and GELU activation [14]. We utilize an Adam optimizer with learning rate 1×10^{-4} and dropout rate of 0.1. We set the number of diffusion steps to $T = 200$. We use a cosine beta schedule [15] and a Gumbel-softmax distribution for noise sampling during the diffusion process. All models are implemented in PyTorch [16] and trained on a Nvidia A6000

GPU with 48GB memory. Our code is publicly available at <https://github.com/sfcurre/DIFFER>.

During inference, we follow the work of Zhong et al. [6] to augment the input SMILES strings. 20 random Root-aligned SMILES strings are generated for each input product and provided to `DIFFER` for sampling. Thus, each model in `DIFFER` outputs 20 random samples for each input product, totaling 160 differently generated reactant SMILES strings. Generated reactant SMILES are canonized and ranked according to rate of occurrence to obtain a final distribution over reactant molecules.

Related work

Retrosynthesis models

Early models for retrosynthesis were primarily rule-based, utilizing known scientific knowledge and synthetic reactions to construct algorithms to find retrosynthetic pathways [17–19]. Over time, these rule-based algorithms gave way to more data-driven approaches, which include template-based, semi-template, and template-free generative models. In template-based approaches, models match features of a target product molecule to known reaction mechanisms in order to rank possible reactants [1, 20], while semi-template methods first predict reaction centers and then complete the resulting synthons (incomplete molecules) to form possible reactants [21, 22]. These methods commonly utilize additional molecular representations such as molecular fingerprints [1, 20] or molecular fragments [23] to aid in decision making, rather than low-level SMILES or graph representations.

In contrast to template-based and semi-template methods, template-free generative models learn to construct new reactants directly from the product structure, commonly using transformers [6, 7, 24] or graph neural networks [24, 25]. Here, generative models are sometimes considered to have the advantage, as they are better able to generalize to reactions that may exist outside the set of reaction templates, at the cost of losing scientific interpretability and trust. For the purpose of this work, we primarily focus on template-free methods operating on SMILES structures, but include comparisons to template-based, semi-template, and non-SMILES based methods as baselines. An overview of models used for comparison is provided in Appendix A.

Early template-free models took advantage of sequential models commonly used in natural language modeling. Liu et al. [5] implement a sequence-to-sequence recurrent encoder-decoder model for retrosynthesis which operates on the SMILES [4] encoding of molecules, and demonstrate model efficacy on-par with rule-based mechanisms for retrosynthesis. Irwin et al. introduce Chemformer [7], a transformer architecture

similar to that of BART [12], a widely-used transformer in NLP for both sequence-to-sequence and discriminative learning tasks. Tu et al. [24] leverage the permutation-invariance of graph neural network encoders to construct a graph-to-sequence model, utilizing additional chemical features extracted by RDKit [26]. Finally, Zhong et al. [6] modify the mechanisms by which SMILES strings are constructed for molecules in order to better represent relationships between products and reactants before they are passed to neural machine translation models. Their root-aligned SMILES methodology is able to significantly improve upon state-of-the-art baselines while applying only pre-existing language models.

An important aspect of recent works which focus on encoding and decoding SMILES representations of molecules is the transformer architecture. Proposed in 2017 for natural language tasks [27], transformer architectures are highly efficient, offering good parallelization while achieving strong performance on sequential input. Recent analysis of transformers have compared their inner workings favorably to graph learning algorithms [28], offering some insight as to why they are able to perform strongly on chemistry-related tasks when applied to SMILES strings. One downside of encoder-decoder transformer models lies in their auto-regressive nature. When predicting SMILES outputs, transformer models are generally required to do so token-by-token, re-using previous outputs as the input for the next prediction in the sequence. This requires a significant amount of planning on the part of the transformer, which is unable to amend previous predictions when predicting later parts of the sequence. Because chemical molecules commonly contain non-sequential structures such as rings and branches, we hypothesize that methods which decode the entire SMILES sequence at once rather than token-by-token may achieve greater success: by allowing the model to attend to both early and later positions in the SMILES sequence during inference, we can better model non-sequential interactions between SMILES structures and reduce the impact of sequential representation on model performance. Thus, we turn toward discrete diffusion models as a possible alternative to auto-regressive decoders.

Categorical diffusion models

Diffusion models are a class of generative model that produce high quality results from uniform random noise. In the forward pass, random noise is iteratively added to a target sample over T steps until the sample is indistinguishable from random noise. In the backward pass, the model is trained to iteratively de-noise the random sample until the target sample is reconstructed [29–31]. Conditional diffusion models [32–34] allow for the inclusion

of conditional constraints and properties during the denoising process (often in the form of textual information), helping guide the model toward a target output with specific parameters or properties.

Traditional diffusion models operate over continuous regimes such as image generation [29, 30], hindering their application on discrete problems such as language or molecular generation tasks. To extend these models to discrete domains, many adaptations have been proposed, such as multinomial diffusion [8] and DiffuSeq [9, 10]. DiffuSeq adapts conditional diffusion models for the task of dialogue generation by employing traditional diffusion within the token embedding space. The authors concatenate embedded representations of the target embeddings and the source embeddings, which act as the conditional. Gaussian noise is then applied to the portion of the sample containing the target embeddings, and the source embeddings are left untouched during the diffusion process. Yuan et al. [35] follow the work of DiffuSeq, but utilize alternative network architectures and support adaptive noise scheduling. Dieleman et al. [36] similarly add noise after input embeddings are applied, and jointly learn both the diffusion and embedding model. In contrast, the multinomial diffusion proposed by Hoogeboom et al. [8] applies noising in the discrete domain, repeatedly sampling from a categorical distribution conditioned on the current diffusion timestep and the current sample. This approach is similarly employed by Austin et al. [37] and DiffusionBERT [38]. DiffusionBERT takes an approach similar to MaskPredict [11], a parallel-decoding schema for text generation that operates similarly to diffusion models. Instead of sampling tokens from a categorical distribution when adding noise, MaskPredict and DiffusionBERT randomly mask-out tokens at varying rates during training, which are un-masked by the model. During inference, both methods generate sequences from a sequence of masked tokens, with no need to initialize a noise vector, which may limit diversity in the generated output.

Results

Overall performance

We compare *DiffER* to a variety of template-based, semi-template, and template-free methods for retrosynthesis prediction with unknown reaction types. We report top- k accuracy for $k \in \{1, 3, 5, 10\}$: following the ranking of generated reactants in order of occurrence, the top- k accuracy is reported as the percentage of reactions in which the ground-truth, patented reactant appears in the top- k generated reactants. We construct *DiffER* as an ensemble of eight diffusion models with various hyperparameters and sample from each model

twenty times with augmented input SMILES according to the work of Zhong et al. [6]. Predicted reactants sampled from the diffusion models are ranked according to the number of times they were predicted by the diffusion models. By using multiple different models with different hyperparameters to predict possible reactants for the same reaction, we encourage diversity and stability

Table 1 Top-K accuracy for template-based, semi-template, and template-free retrosynthesis models

Category	Model	K=1	3	5	10
Template-based	Retrosim [20]	37.3	54.7	63.3	74.1
	Neuralsym [1]	44.4	65.3	72.4	78.9
	GLN [39]	<u>52.5</u>	<u>69.0</u>	<u>75.6</u>	<u>83.7</u>
	LocalRetro [23]	53.4	77.5	85.9	92.4
Semi-template	G2Gs [40]	48.9	67.6	72.5	75.5
	GraphRetro [22]	53.7	68.3	72.2	75.5
	RetroXpert [41]	50.4	61.1	62.3	63.4
	RetroPrime [42]	51.4	70.8	74.0	<u>76.1</u>
	G ² Retro [21]	<u>53.9</u>	<u>74.6</u>	<u>80.7</u>	86.6
	GDiffRetro [43]	58.9	79.1	81.9	-
Template-free	Seq2Seq [5]	37.4	52.4	57.0	61.7
	Levenshtein [44]	41.5	48.1	50.0	51.4
	GTA [45]	51.1	67.6	74.8	81.6
	Graph2SMILES [24]	51.2	66.3	70.4	73.9
	Dual-TF [46]	53.3	69.7	73.0	75.0
	MEGAN [25]	48.1	70.7	78.4	86.1
	Chemformer [7]	54.3	-	62.3	63.0
	Retroformer [47]	53.2	71.1	76.6	82.1
	Tied transformer [48]	47.1	67.2	73.5	78.5
	R-SMILES [6]	<u>56.3</u>	79.2	86.2	91.0
	DiffER	57.6	<u>79.0</u>	<u>84.1</u>	<u>87.4</u>

The best performing model of its category for each K is in **bold** and the second best model is underlined. DiffER is the best performing model for $K = 1$ and the second best model for $K = 3, 5, 10$ among template-free methods

in the output reactants. Ties in reactant rankings are decided according to a ranked choice voting scheme. Further ties are broken arbitrarily. Results for the diffusion ensemble model are reported against baseline algorithms in Table 1, while individual model results are reported in Table 3. We include a short review of the baselines used for comparison in appendix A.

Notably, DiffER is the best performing model for $K = 1$ (57.6%) across all template-free models and is highly competitive for higher values of K , achieving the second best performance for $K = 3$, $K = 5$, and $K = 10$ (79.0%, 84.1%, and 87.4%, respectively) for template-free methods behind Root-Aligned SMILES. Comparing to the state-of-the-art $K = 1$ result of GDiffRetro [43], DiffER achieves comparable $K = 3$ performance and out-performs GDiffRetro at higher values of K , indicating DiffER produces a wider array of possible reactants. Notably, compared to R-SMILES, both DiffER and GDiffRetro struggle with output diversity, resulting in both diffusion-based models having decreased performance at high K . The decrease in performance for these models at high K can likely be attributed to the affinity for diffusion models to most often sample from the peak of the posterior distribution: despite generating outputs from multiple models for a diversity of input product SMILES, the output of DiffER tends to produce just a few candidate reactant sets for each product, producing the same output molecule numerous times in a different, non-canonical SMILES form. Indeed, Fig. 2 shows a histogram of the number of sample reactions generated for the test dataset. The diffusion ensemble produces a median number of samples of 5, with 17.9% of reactions producing less than 5 samples, and 57.1% producing less than 10. Notably, we find that fewer reactants tend to be sampled when DiffER is correct in its output: reactions where the top-1 reactant is correctly predicted produce

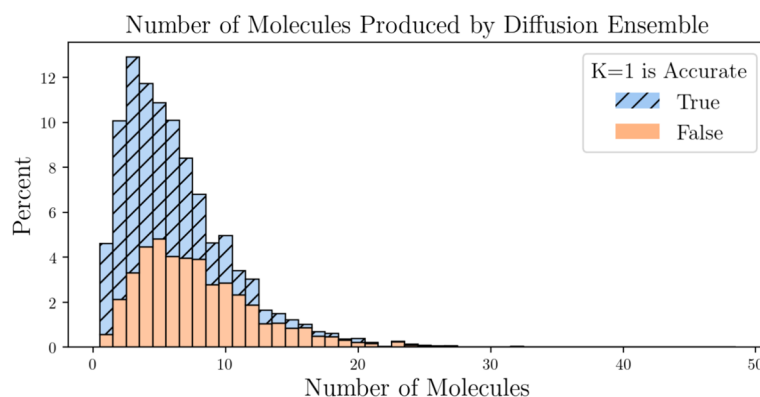


Fig. 2 A histogram of the number of different molecules output by DiffER for each reaction during inference. The distribution has a mean of 10.0 and a median of 9. “K=1 is Accurate” indicates if the most commonly predicted molecule matched the ground-truth reaction

only 8.4 different molecules on average, compared to 12.3 for cases where the top-1 reactant does not match the ground truth. This lack of outputs is less observed in other methods such as R-SMILES which apply beam-search algorithms to sequential decoders, allowing a greater number of possible reactants to be found.

We further analyze the performance of *DiffER* across varying reaction types, as presented in Table 2. Reaction types are determined by Schneider et al. [49]. Notably, the performance of *DiffER* is inconsistent among reaction types, which has also been observed in other models [21]. The diffusion ensemble performs strongly on reactions which operate on ring structures, such as acylation, but performs poorly at predicting patented reactants for ring formation, through either heterocycle formation

or C-C bond formation. Compared to *G²Retro* [21], the best performing semi-template method which offers a similar analysis of their model, *DiffER* is significantly more capable of accurately predicting the reactants for protection reactions [21]. For other reaction types, the overall trends in performance are rather similar for both *DiffER* and *G²Retro*. Both models exhibit their best performance on functional group addition followed by acylation, and their worst performance on heterocycle formation, functional group interconversion, and C-C bond formation.

Individual model performance

We construct *DiffER* from eight individually trained diffusion models, each of which uses a different random padding limit N (described in Sect. Length prediction) during training. We report the individual performance for each of the trained models in Table 3. We compare these results to the *DiffER* ensemble as well as a baseline diffusion model trained using direct length prediction, without the length variation used in the individual *DiffER* models, also reported in Table 3.

The individual *DiffER* models greatly improve upon the baseline length prediction model, demonstrating the effectiveness of our length variation technique. The length variation technique encourages the individual models to over-predict the length of the reactant SMILES and then reduce the output SMILES to the correct size using predictable padding tokens. Meanwhile, the baseline length prediction model drastically under-performs all other models in terms of accuracy, demonstrating the sensitivity of the diffusion models for SMILES string

Table 2 Top-K accuracy for *DiffER* grouped by reaction type

Reaction Type	K=1	3	5	10	Support
Heteroatom alkylation and arylation	58.6	81.1	86.0	89.3	1,516
Acylation and related processes	70.4	88.7	92.2	93.9	1,190
Deprotections	53.4	79.8	84.9	87.2	822
C-C bond formation	42.2	64.2	72.0	77.8	567
Reductions	57.2	75.2	81.4	86.2	463
Functional group interconversion	39.0	58.2	66.5	73.1	182
Heterocycle formation	33.3	48.9	55.6	60.0	90
Oxidations	60.2	84.3	88.0	91.6	83
Protections	64.2	83.6	91.0	91.0	67
Functional group addition	78.3	87.0	87.0	87.0	23

Rows are ordered by number of reactions of the specified reaction type in descending order

Table 3 Comparison of the *DiffER* ensemble with the individual models that make up *DiffER*

Model	K=1	3	5	10	Sample Validity	Avg. Num. Reactants
<i>DiffER</i> Ensemble	57.6	79.0	84.1	87.4	100.0	10.0
<i>DiffER</i> ₂₀	53.2	70.3	72.9	73.6	100.0	3.2
<i>DiffER</i> ₃₀	55.2	71.5	74.4	75.2	99.9	3.2
<i>DiffER</i> ₄₀	54.6	72.1	74.4	75.3	99.9	3.2
<i>DiffER</i> ₅₀	54.9	71.3	73.7	74.4	100.0	3.2
<i>DiffER</i> ₆₀	55.4	71.7	74.6	75.4	99.9	3.3
<i>DiffER</i> ₇₀	54.3	71.1	73.5	74.4	99.6	3.2
<i>DiffER</i> ₈₀	54.6	71.9	74.4	75.1	99.6	3.3
<i>DiffER</i> ₉₀	54.5	71.6	74.2	74.9	99.8	3.3
Baseline Length	40.4	55.9	58.8	59.9	99.9	3.8
Oracle Length	77.0	88.1	89.5	90.0	99.7	2.8

Individual models are notated as *DiffER* _{N} , where N represents the upper bound on the number of added padding tokens according to the procedure in Sect. Length prediction. We additionally include single models with baseline length prediction (length prediction without padding) and oracle length prediction (perfect length prediction without padding). These models are used to show the efficacy of the variant padding approach and the upper limit of diffusion model performance with perfect length prediction, respectively. The individual *DiffER* models show significantly higher accuracy than baseline length prediction, but are outperformed by the *DiffER* ensemble. The oracle length model drastically outperforms all existing models

generation to the size of the noise vector. Both the individual DiffER models and the baseline length prediction model exhibit low variance in the output reactants, only predicting 3.2 and 3.8 reactants on average, respectively. Notably, all models still achieve high validity from among the output molecules, indicating that the individual diffusion models are properly learning not only the SMILES grammar, but also rules of molecular validity.

When the individual models are ensembled into the complete DiffER model, both the accuracy and diversity of predicted reactants are significantly improved. The DiffER ensemble produces on average three times the number of reactants compared to the individual models, highlighting the impact of the different individual model setups on the output molecules. By combining multiple models trained on different upper limits for the length of the SMILES string, we can leverage different weightings of the length prediction mechanism. When a model is trained on a lower padding limit, more emphasis is placed on the length prediction component, as the model is less able to diverge from the predicted length. When the padding limit is larger, the diffusion model has more freedom to construct SMILES of varying lengths, but are less informed by the length prediction. By combining multiple models on various random padding limits, we improve the diversity and accuracy of the predicted reactants. These results highlight the significant performance improvement obtained by ensembling diffusion models. While individual models have limited top-k performance compared to auto-regressive models, the increased diversity and capacity of the ensemble method leads to notable improvement in model performance, resulting in the observed state-of-the-art top-1 accuracy and competitive top-3, 5, and 10 accuracy among template-free methods.

Upper Limits on Performance

In this section, we consider an upper limit on the performance of categorical diffusion models for chemical retrosynthesis by considering an oracle model which predicts the change in length between product and reactant SMILES with 100% accuracy. This allows us to construct a diffusion model without needing to incorporate aspects of length prediction or variability in the size of the output as discussed in Sect. [Length prediction](#). Under the assumption of the length-predicting oracle, we can initialize input noise to the diffusion model of the exact reactant size. We run this experiment for a single diffusion model and utilize the same parameters and repeated sampling during inference as discussed in the experimental setup.

Results for this experiment are presented in Table 3 in comparison to the DiffER models and the baseline length prediction diffusion model. The model with

oracle length drastically outperforms DiffER as well as all existing methods for $K = 1, 3, 5$, but performs slightly worse than both Root-aligned SMILES and LocalRetro for $K = 10$. Similarly to DiffER, this can be attributed to a lack of variety in the predicted output as a result of oversampling from posterior peaks, which is even more extreme for the oracle length model: on average, only 2.8 different reactants are predicted, with a median of just 2 predicted reactants. However, this lack of diversity is overcome by the oracle model's high accuracy among the predicted reactants.

This result highlights the importance of accurate length prediction in non-autoregressive models, particularly in the case of molecular generation with SMILES strings. While there is some variability in the length of randomly generated SMILES string for a specific molecule, these variants generally only differ by a length of two or three, if they differ at all. If the length prediction is off from a viable value by even one or two, it may force an entirely different molecule to be generated, resulting in lower reported performance. By including the random length padding in DiffER, we help combat the diffusion model's sensitivity to length prediction, but remain far from the performance of a model with perfect length prediction.

Potential Impact of Accurate Length Prediction

The performance of the oracle length prediction model demonstrates the impact successful length prediction could have on the diffusion model sampling procedure. If we are able to develop methods to effectively query or predict reactions with specific changes in sequence length, we can significantly improve the model's efficacy. From a chemical perspective, the difference in length between product and reactant SMILES can be viewed as a proxy for determining the size and structure of a leaving group: SMILES sequences with longer length tend to contain either more atoms or larger amounts of branching in the molecular structure. Thus, smaller changes in length limit the model to performing specific reactions with smaller leaving groups, such as acylation reactions with acid halides, which most commonly result in the removal of only a single halide atom from the SMILES string when transitioning from reactant to product. In contrast, larger changes in length restrict the model toward larger leaving groups. A commonly observed example in the USPTO-50K dataset [13] are C-C bond forming reactions via Suzuki cross-coupling with boronic esters [50]. These reactions commonly have a 1,3,2-dioxaborolane leaving group, which results in a length change > 18 (the length of the 1,3,2-dioxaborolane SMILES string) between product and reactant SMILES. Cross-coupling reactions with 1,3,2-dioxaborolane comprise only 2.3% of the test set, but encompass nearly 40%

of all reactions in the test set with a change in SMILES length > 18, and 94% of reactions with a specific length change of 24. Thus, if the diffusion model is prepared with a larger change in sequence length, the likelihood of producing an accurate prediction of a cross-coupling reaction with 1,3,2-dioxaborolane can be significantly improved. By accurately predicting specific changes in sequence length, we can significantly reduce the space of possible reactants and drastically improve the sensitivity of the model toward specific reactions.

Case studies

Our final results are in the form of a case study of four different reactions, the results of which can be seen in Fig. 3. We select four different reactions from the test dataset to highlight the successes and limitations of the diffusion models. For each reaction, we report the top-3 reactants sampled from the diffusion ensemble.

Reaction **a** (Fig. 3a) is a deprotection reaction in the forward direction in which the trifluoroacetamide protecting group (Fig. 3ab) is reacted with sodium borohydride in an ethanol solution to produce a primary amine at the N1 position (Fig. 3aa). The diffusion ensemble does not predict this reactant, however, instead opting for the reduction of a nitro group, predicted 89.4% of the time, (Fig. 3ac) at the N1 position. This is a widely used and viable reaction involving metal catalyzed reduction of the nitro group using Fe, Zn, Pd or Ni. Similarly, the third ranked choice (Fig. 3ae) is also a viable option involving a debenzyl reaction which is commonly performed in the presence of a metal catalyst under a hydrogen atmosphere. The second choice reaction (Fig. 3ad), however, undergoes amide hydrolysis under harsh conditions such as a strong acid. While this is also a viable synthesis, a harsh condition such as a strong acid would likely interact with the tert-butyl ester group leading to unwanted side products. Notably, the diffusion ensemble does not predict the patented reactant, primarily predicting the reduction of a nitro group 89.4% of the time. We hypothesize that this is a result of the specific use-case of the ground-truth reaction, which is likely used in an industrial setting. Unlike the top-1 (Fig. 3ac) reactant, the ground-truth (Fig. 3ab) does not use the metal catalyzed reduction of a nitro group (top-1) or debenzyl reaction (top-3) under a hydrogen atmosphere because it is not a safe reaction on industrial scale. While the top-1 or top-3 reactants may actually be more common in an academic lab setting, the ground-truth reactant is likely preferable on a large-scale industrial setting.

Reaction **b** (Fig. 3b) is a common reduction reaction of a ketone at C1 (Fig. 3bb) with sodium borohydride in a methanol solution to produce a secondary alcohol (Fig. 3ba). This same reaction is also predicted by the

diffusion models 38.3% of the time to give the second most popular predicted reactants (Fig. 3bd). The most popular predicted reactants is an aldehyde and Grignard reagent, predicted 42.2% of the time, in which the methylmagnesium bromide reacts with the aldehyde at position C1 (Fig. 3bc) to produce the desired product in the forward reaction. However, the Grignard reagent (MeMgBr), which is a more nucleophilic reagent, is likely to react with other functional groups present in the molecule as well, such as the two amides at C2 and C4 or displace the fluorine at C5 via a nucleophilic aromatic substitution-like type reaction. In contrast, while the ground-truth/second-most predicted reaction (Fig. 3bb/3bd) could lead to unwanted products, such as the reduction the carbonyl groups at C2 and C4, these reactions are less likely to occur compared to the Grignard reaction under a proper choice of reagents such as sodium borohydride in ethanol. The top-3 reactant is predicted just 9.7% of the time, and unlike the other proposed reactants, it does not interact with the C1 carbon. Instead, it disconnects the molecule at the C1-N3 amide to reveal a C1 carboxylic acid and an N3 amine. (Fig. 3be). The amide bond can be constructed in the forward reaction using a suitable amide coupling reaction condition. Yet again, this reaction is viable but would likely produce unwanted byproducts if other functional groups such as the free secondary alcohol or the primary amide in the proposed reaction were not protected first.

Reaction **c** (Fig. 3c) is an ester hydrolysis in which the C1 ester (Fig. 3cb) is reacted with sodium hydroxide and a water/ethanol mixture to produce a carboxylic acid (Fig. 3ca). This reaction is also predicted by the diffusion ensemble a vast majority of the time, forming 61.3% of the sampled reactants (Fig. 3cc). The top-2 reactant is also a viable alternative to the ground-truth reactant, just with a ethyl ester at C1 instead of a methyl ester, which is also a common reaction. Interestingly, the top-3 reactant sampled by the diffusion models is a Paal-Knorr-type pyrrole synthesis [51, 52] from an aniline (N2) and the diketone (C3-C6)(Fig. 3cd). The Paal-Knorr reactions usually involves harsh conditions such as strong acids which may not be compatible with the ether or the lactam functional groups [51, 52]. Also, the free C1 carboxylic acid may need to be protected as an ester before the pyrrole synthesis to produce the ground-truth reactant molecule (Fig. 3cb) which would need to undergo the aforementioned ester hydrolysis to produce the target molecule. Thus, 3cd could be viewed as one-step further back in the retrosynthesis chain compared to the ground-truth/top-1 reactant.

Finally, reaction **d** (Fig. 3d) is a Borch reductive amination [53] between cyclobutanone and a secondary amine in the presence of a reducing agent (Fig. 3db) to form a

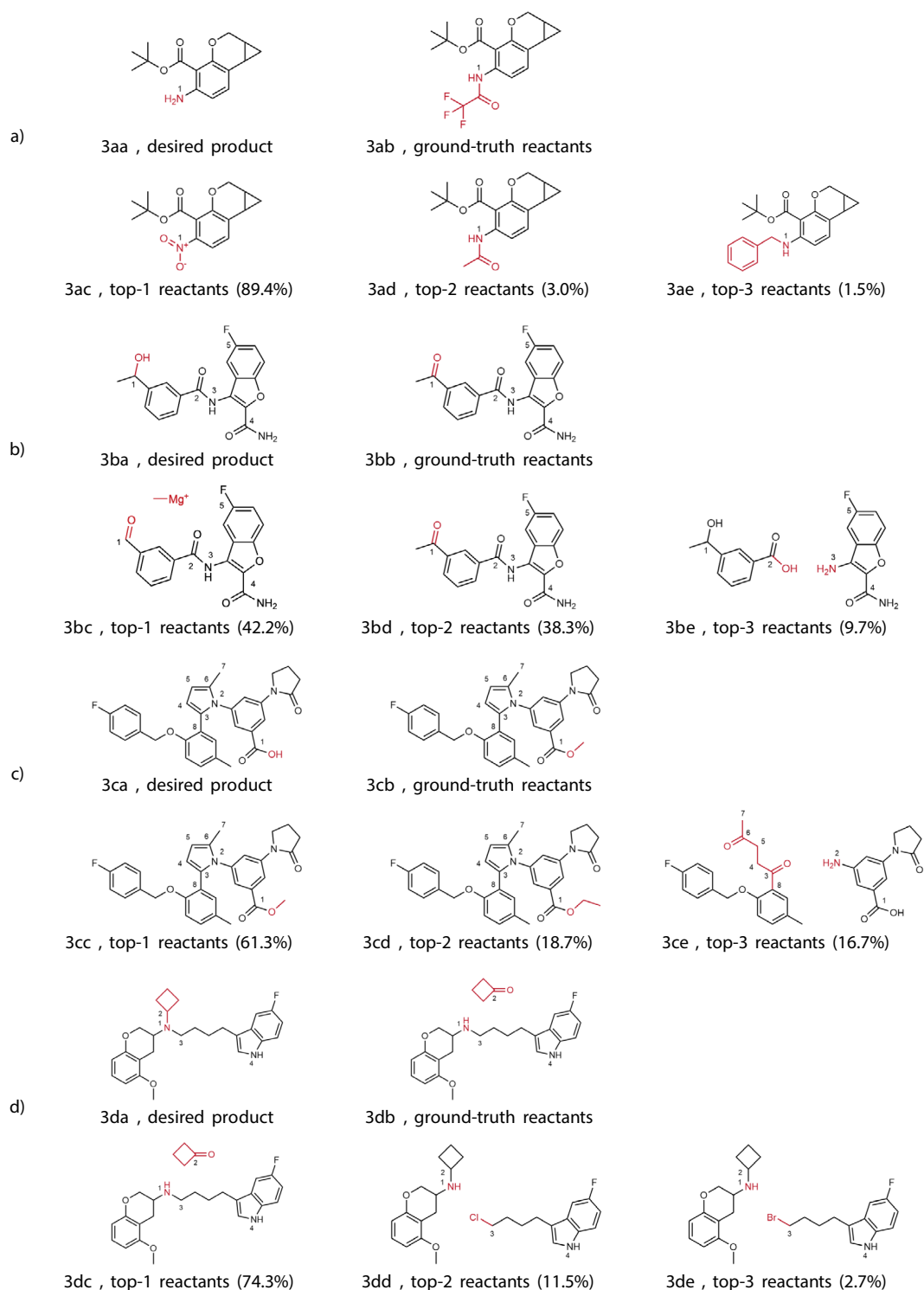


Fig. 3 Top-3 reactants for 4 different target products from the test set compared to patent reactants. Values in parentheses indicated the percentage of time that reactant was sampled from the diffusion models. Numbers next to atoms are labels used to refer to the atom. Colored regions indicate where in the molecule a reaction is expected to occur

tertiary amine with a cyclobutyl group at N1 (Fig. 3db). The ground-truth reactants are the top-predicted reactants by *DiffER*, making up 74.3% of the predicted reactants. Both the second and third most predicted reactants are amine alkylations involving halides, in which an alkyl halide at C3 (Fig. 3dd/3de) reacts with the N1 amine to form a tertiary amine. Such N-alkylation reactions are commonly used when reduction amination does not work, and are valid alternative reactions. However, potential side reactions such as N-1 overalkylation or intramolecular N-4 alkylation would likely lead to unwanted byproducts. As such, the ground-truth/top-1 reactants are preferred.

These case studies demonstrate that the diffusion models are capable of learning a variety of viable reactions for organic synthesis, but like many other models, struggle to learn greater domain requirements or the likelihood of byproducts, as exemplified in the top-1 reactant for reaction **b** (Fig. 3bc) and top-2/3 reactants for reaction **d** (Fig. 3dd/3de). However, the models also demonstrate significant capability to properly utilize common and routine reactions (Fig. 3ac and Fig. 3cc) as well as more complex reactions such as ring formations (Fig. 3ce). Even in cases where the ground-truth reactants were not predicted in the top-3 reactants, the diffusion models produced suitable alternatives used in other reactions (Fig. 3ac) to accomplish the same goal. This highlights a key difficulty in assessing the accuracy and performance of models for retrosynthesis; oftentimes, multiple different reactions may be viable, and the use case for each reaction may depend on information not present in the molecular structure, such as the case of reaction **a**. Metrics of ground-truth accuracy do not capture the viability of a reaction, just that they match the patented reaction, which may lead to misrepresentation of the model's capabilities when a suitable alternative is predicted.

Discussions and conclusions

Our results show that categorical diffusion models offer a competitive alternative to auto-regressive models for template-free single step retrosynthesis prediction and are able to outperform many state-of-the-art template-based, semi-template, and template-free methods on top-1 accuracy, and achieve similar performance to state-of-the-art models on top-3, top-5, and top-10 accuracy. The proposed *DiffER* ensemble is able to efficiently sample from the posterior distribution of possible reactant SMILES strings while maintaining the viability of output SMILES. *DiffER* is capable of reproducing existing patented reactions as well as proposing new and viable reactants, demonstrating the diffusion models' capability to effectively learn and apply common reaction patterns as well as distill necessary reactivity information

from SMILES strings. We hope that *DiffER* provides new directions of research into developing statistical machine learning models that may aid domain experts in retrosynthesis planning.

Comparison to other template-free methods

Compared to other template-free methods, our diffusion ensemble is able to construct the entire reactant SMILES string in unison, rather than token by token. Methods like Chemformer [7] or R-SMILES [6] build the SMILES sequence by predicting a distribution of probabilities for the next token conditioned on the current sequence and then selecting the token with the highest probability and adding it to the sequence. To generate multiple possible reactants, auto-regressive methods generally use beam-search algorithms to generate multiple possible reactants [7]. Beam-search algorithms explore multiple options at each step, and only continue exploring the top-k sequences with the highest probability. In contrast, our diffusion approach directly samples an entire reactant SMILES string from the approximated posterior distribution of possible reactants. Thus, the proportion of times a molecule is sampled from *DiffER* can be interpreted as a approximation of the posterior probability of that reactant. This offers an interpretable measure of confidence in the model outputs. However, this form of sampling also results in limited diversity in the output reactant set: when the model is highly confident or little diversity exists in the space of possible reactants, only a few possible reactants may be produced. In contrast, when the model is unsure or there is a high number of possible reactions, more reactants will be produced. This trade-off can be disadvantageous in certain circumstances, particularly when there is a low number of predicted reactants and the top-predicted reactant is unsuitable.

Comparison to template-based and semi-template methods

Unlike template-based and semi-template methods, diffusion models do not directly utilize a static set of reaction templates. Rather, *DiffER* learn suitable reactions directly from the data, allowing for greater generalization to reactions that may occur outside the realm of existing templates. However, this comes at a notable cost in interpretability: because the diffusion method constructs the proposed reactants from randomly generated uniform noise, it is not always clear what causes the emergence of a particular reactant over another possible reactant. A greater analysis of the latent noise would need to be done to further understand the workings of *DiffER*. In contrast, template-based and semi-template based methods directly utilize the existing molecular structure and set of reaction templates when determining where

reaction centers are and which reaction templates apply. This offers a greater level of interpretability and scientific backing than template-free methods.

Limitations and directions of future work

Despite the strong performance of *DiffER*, there are a few notable limitations to the modeling approach. Firstly, raw diffusion models are highly sensitive to the predicted size of the output. Thus, methods such as our novel variant length padding (Sect. [Ensemble voting of models](#)) must be employed to diversify target lengths and allow diffusion models to predict molecules of varying size. Secondly, *DiffER* suffers from a lack of output variety, often producing just a few possible reactants. Because diffusion models sample from an approximated posterior, reactants which the model deems more probable will be sampled at a higher rate, resulting in fewer reactants being sampled overall for a static sample size. This is particularly true in cases where the model is confident in its prediction. Finally, the model sometimes produces sub-optimal reactions, such as some of those displayed in Fig. 3. Such reactions often undergo additional reactions alongside the target reaction, and while the desired product is a possible outcome, additional byproducts would likely be produced in many cases. These limitations highlight necessary areas of future research for not only categorical diffusion models in chemical retrosynthesis, but also categorical diffusion models for sequence generation in general:

1. Categorical diffusion models must be able to adapt to differently sized outputs. Prior work relies on predictable padding tokens to predict differently sized sequences [9, 10], but this reliance can introduce its own biases due to the prevalence of padding tokens for shorter sequences. *DiffER* overcomes reliance on predicting padding tokens by adding limited variability in the number of possible padding tokens, with some success; however, results still pale in comparison to a model with perfect sequence length prediction.
2. Additional techniques to sample categorical diffusion models must be developed in order to improve sequence diversity and coverage. Compared to the traditional applications of diffusion models in image generation, molecular SMILES generation has significantly fewer possible outputs, making the models more likely to converge to a single prediction. This effect is heightened by the the presence of multiple SMILES strings mapping to the same molecular structure: while the sequence itself may be different, the same molecule is produced. We must develop methods to encourage greater output diversity in the

sampled molecules. *DiffER* attempts encourages output diversity using an ensemble approach, and while the results are significantly more diverse than single model approaches, the number of differently sampled molecules remains low.

The case studies of *DiffER* additionally demonstrate directions of future research for retrosynthetic modeling as a whole, many of which are echoed in other work [21]:

3. Machine learning models for retrosynthesis are generally unaware of possible side products as a result of the generally available training data. Because these models are most commonly trained on patent reaction datasets, models are not generally exposed to explicit examples where multiple possible outcomes could occur. Indeed, the UPSTO-50K dataset has no reactant sets which map to more than one molecule, limiting the ability of the model to learn that multiple reactions may be possible for a single set of reactants. To improve the ability of ML models to learn chemical processes, incorporating additional examples into the training datasets may be beneficial, and with proper learning techniques, may lead to improved performance.
4. Evaluating the performance of ML models on chemical retrosynthesis most commonly relies on a) matching to patented reactions and b) analyzing individual case studies. Evaluating models by matching to patented reactions limits the discovery of novel techniques and pathways, as models may produce valid results that do not match patent data. In contrast, evaluating individual case studies is more adaptable to understanding the model's true performance, but is a time consuming process requiring considerable chemical knowledge. To enhance the development of ML models for chemical retrosynthesis, we must develop new methods of analysis that take into account the abundance of viable reactions as well as existing chemical knowledge.

Further research in these four tasks would greatly benefit the application of diffusion models in chemical retrosynthesis, as well as both diffusion modeling and retrosynthetic prediction models individually.

Conclusion

Our ensemble of categorical diffusion models provides state-of-the-art top-1 accuracy as well as competitive top-3, top-5, and top-10 accuracy for template-free methods. We demonstrate that our ensemble model is capable of not only reproducing patented reactions, but also producing viable alternative reactions for the same

product. Additionally, we offer insight into the importance of length prediction and length variation methods when training diffusion models for sequence prediction tasks. Finally, we include an analysis and discussion of the limitations of our approach. In doing so, we hope to provide a suitable baseline and open new avenues of exploration for a new class of template-free retrosynthesis model which differs from current autoregressive approaches. In future work, we plan to investigate additional methods of length prediction, as well as apply different diffusion sampling techniques to improve the diversity of sampled molecules. Finally, we plan to continue improving upon the diffusion model that comprise *DiffER*, introducing methods such as adaptive noise scheduling [35] as well as other advancements in categorical diffusion models to improve model performance.

Overview of Retrosynthesis Models Used for Comparison

We compare and report results against a variety of retrosynthesis models, as seen in Table 1. Models are grouped according to their model type, which includes template-based, semi-template, and template-free models.

The template-based models we compare against are:

- Retrosim [20], which applies reaction templates derived from molecules similar to the target product, and ranks reactions by similarity to existing reactions;
- Neursym [1], which utilizes multi-layer perceptrons with molecular fingerprints of the product to predict applicable reaction templates;
- GLN [39], which uses graph neural networks to learn when and where in molecules reaction rules can be applied, while scoring the feasibility of the reaction;
- LocalRetro [23], which focuses on the local environment of atoms in the product and classifies reaction templates on an atomic level.

Each of these methods utilize reaction templates sourced from known reactions. The primary goals are to 1.) select where in the molecule the reaction will take place and 2.) select which reaction template is applicable.

The semi-template models we compare against are:

- G2Gs [40], which predicts reactions centers and sequentially completes synthons through the use of a variational graph autoencoder;
- GraphRetro [22], which uses message-passing neural networks to classify reaction centers and completes

synthons by predicting which leaving groups to add from among a pre-selected set;

- RetroXpert [41], which uses edge-enhanced graph attention networks to predict reaction centers and completes synthons using a transformer network;
- RetroPrime [42], which uses two separate transformers, one to predict reaction centers, and the other to map synthons to reactants;
- G²Retro [21], which uses a message passing network to predict three different types of reaction centers and sequentially adds substructures on to the synthons until complete reactants are formed.
- GDiffRetro [43], which uses a dual graph molecular representation to train a reaction center prediction model and uses a conditional diffusion model to complete the produced synthons.

Each of these methods first divide the product into synthons and then transform the synthons into complete reactants. The primary goals are to 1.) select where in the molecule the reaction will take place and 2.) complete synthons formed from the divided product into complete molecules which will react as needed.

The template-free models we compare against are:

- Seq2Seq [5], which applies sequence-to-sequence encoder-decoder recurrent neural networks to predict reactant SMILES from the product SMILES;
- Levenshtein [44], which augments the training datasets of sequence-to-sequence recurrent neural networks by ensuring that the source and target SMILES strings have similar subsequences;
- GTA [45], which incorporates graphical information in a sequence-to-sequence model by limiting the self-attention layer using the adjacency matrix of the product molecular graph;
- Graph2SMILES [24], which leverages the permutation invariant nature of graph structures to remove variance that occurs when product is represented in SMILES form, and leverages transformer architectures to predict the reactant SMILES;
- Dual-TF [46], which unifies graph-based and sequence-based methods to learn an energy-based model which ranks possible reactants according to their energy score;
- MEGAN [25], which models one-step retrosynthesis as a series of graph edits, and trains an encoder-decoder graph attention model to learn how to adapt the product into a set of reactants;
- Chemformer [7], which pretrains a transformer architecture on SMILES strings using a masking approach and fine-tunes the model on product and reactant strings for retrosynthesis;

- Retroformer [47], which jointly processes the molecular sequence and graph and uses localized attention to relay information between the reaction center and global chemical context when constructing reactants;
- Tied transformer [48], which uses two-way transformers to encourage diversity and grammatical accuracy in predicted SMILES strings;
- R-SMILES [6], which restructures SMILES representations of products and reactants to have significant structural overlap, and then uses neural machine translation architectures to map from products to reactants.

Most of these methods utilize transformer architectures on SMILES strings, effectively translating from a product string into a reactant string. Many methods also directly incorporate information from the molecular graph, augmenting the SMILES strings with known structural information. The primary goal is to directly predict a representation of the reactant from the product, while secondary goals are to ensure the validity and viability of the predicted reactant representation.

Author contributions

SC designed, conceptualized, and implemented the methodology and wrote the manuscript with input from XN and SP. ZC assisted in reviewing the methodology and code. DAA assisted in the case study. All authors helped write the manuscript. All authors read and approved the final manuscript.

Funding

This project was made possible, in part, by support from the National Science Foundation grant no. IIS-2133650 and the National Library of Medicine grant no. 1R01LM014385. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agency.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no Competing interests.

Received: 23 January 2025 Accepted: 12 July 2025

Published online: 29 July 2025

References

- Segler MH, Waller MP (2017) Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry-A Eur J* 23(25):5966–5971
- Segler MH, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic ai. *Nature* 555(7698):604–610
- Shen Y, Borowski JE, Hardy MA, Sarpong R, Doyle AG, Cernak T (2021) Automation and computer-assisted planning for chemical synthesis. *Nature Rev Methods Primers* 1(1):1–23
- Weininger D (1988) Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 28(1):31–36
- Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, Ho S, Sloane J, Wender P, Pande V (2017) Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent Sci* 3(10):1103–1113
- Zhong Z, Song J, Feng Z, Liu T, Jia L, Yao S, Wu M, Hou T, Song M (2022) Root-aligned smiles: a tight representation for chemical reaction prediction. *Chem Sci* 13(31):9023–9034
- Irwin R, Dimitriadis S, He J, Bjerrum EJ (2022) Chemformer: a pre-trained transformer for computational chemistry. *Mach Learning Sci Tech* 3(1):015022
- Hoogeboom E, Nielsen D, Jaini P, Forré P, Welling M (2021) Argmax flows and multinomial diffusion: Learning categorical distributions. *Adv Neural Inf Process Syst* 34:12454–12465
- Gong S, Li M, Feng J, Wu Z, Kong L (2023) Diffuseq: Sequence to sequence text generation with diffusion models. In: *International Conference on Learning Representations (ICLR 2023)* (01/05/2023–05/05/2023, Kigali, Rwanda)
- Gong S, Li M, Feng J, Wu Z, Kong L (2023) Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*
- Ghazvininejad M, Levy O, Liu Y, Zettlemoyer L (2019) Mask-predict: Parallel decoding of conditional masked language models. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics
- Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L (2020) Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871. Association for Computational Linguistics
- Lowe D (2017) Chemical reactions from US patents (1976–Sep2016) <https://doi.org/10.6084/m9.figshare.5104873.v1>
- Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*
- Chen T (2023) On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al (2019) Pytorch: An imperative style, high-performance deep learning library
- Cook A, Johnson AP, Law J, Mirzazadeh M, Ravitz O, Simon A (2012) Computer-aided synthesis design: 40 years on. *Wiley Interdisciplinary Rev Comput Molecular Sci* 2(1):79–107
- Corey EJ, Long AK, Rubenstein SD (1985) Computer-assisted analysis in organic synthesis. *Science* 228(4698):408–418
- Sun Y, Sahinidis NV (2022) Computer-aided retrosynthetic design: fundamentals, tools, and outlook. *Curr Opin Chem Eng* 35:100721
- Coley CW, Rogers L, Green WH, Jensen KF (2017) Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent Sci* 3(12):1237–1245
- Chen Z, Ayinde OR, Fuchs JR, Sun H, Ning X (2023) G 2 retro as a two-step graph generative models for retrosynthesis prediction. *Commun Chem* 6(1):102
- Somnath VR, Bunne C, Coley C, Krause A, Barzilay R (2021) Learning graph models for retrosynthesis prediction. *Adv Neural Inf Process Syst* 34:9405–9415
- Chen S, Jung Y (2021) Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* 1(10):1612–1620
- Tu Z, Coley CW (2022) Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *J Chem Inf Model* 62(15):3503–3513
- Sacha M, Błaz M, Byrski P, Dabrowski-Tumanski P, Chrominski M, Loska R, Włodarczyk-Pruszyński P, Jastrzebski S (2021) Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *J Chem Inf Model* 61(7):3273–3284
- Landrum G et al (2013) Rdkit: a software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 8(31.10):5281
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Adv Neural Inf Process Syst* 30:2

28. Joshi CK (2020) Transformers are Graph Neural Networks. Towards Data Science. <https://towardsdatascience.com/transformers-are-graph-neural-networks-bca9f75412aa>
29. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR
30. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 33:6840–6851
31. Nichol AQ, Dhariwal P (2021) Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171. PMLR
32. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, Ghasemipour K, Gontijo Lopes R, Karagol Ayan B, Salimans T et al (2022) Photorealistic text-to-image diffusion models with deep language understanding. *Adv Neural Inf Process Syst* 35:36479–36494
33. Gu S, Chen D, Bao J, Wen F, Zhang B, Chen D, Yuan L, Guo B (2022) 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
34. Zhang L, Rao A, Agrawala M (2023) Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3836–3847
35. Yuan H, Yuan Z, Tan C, Huang F, Huang S (2024) Text diffusion model with encoder-decoder transformers for sequence-to-sequence generation. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 22–39
36. Dieleman S, Sartran L, Roshannai A, Savinov N, Ganin Y, Richmond PH, Doucet A, Strudel R, Dyer C, Durkan C et al (2022) Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*
37. Austin J, Johnson DD, Ho J, Tarlow D, Van Den Berg R (2021) Structured denoising diffusion models in discrete state-spaces. *Adv Neural Inf Process Syst* 34:17981–17993
38. He Z, Sun T, Tang Q, Wang K, Huang X-J, Qiu X (2023) Diffusionbert: Improving generative masked language models with diffusion models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4521–4534
39. Dai H, Li C, Coley C, Dai B, Song L (2019) Retrosynthesis prediction with conditional graph logic network. *Adv Neural Inf Process Syst* 32:2
40. Shi C, Xu M, Guo H, Zhang M, Tang J (2020) A graph to graphs framework for retrosynthesis prediction. In: International Conference on Machine Learning, pp. 8818–8827. PMLR
41. Yan C, Ding Q, Zhao P, Zheng S, Yang J, Yu Y, Huang J (2020) Retroxpert: decompose retrosynthesis prediction like a chemist. *Adv Neural Inf Process Syst* 33:11248–11258
42. Wang X, Li Y, Qiu J, Chen G, Liu H, Liao B, Hsieh C-Y, Yao X (2021) Retroprime: a diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem Eng J* 420:129845
43. Sun S, Yu W, Ren Y, Du W, Liu L, Zhang X, Hu Y, Ma C (2025) Gdiffretro: Retrosynthesis prediction with dual graph enhanced molecular representation and diffusion generation. *arXiv preprint arXiv:2501.08001*
44. Sumner D, He J, Thakkar A, Engkvist O, Bjerrum EJ (2020) Levenshtein augmentation improves performance of smiles based deep-learning synthesis prediction
45. Seo S-W, Song YY, Yang JY, Bae S, Lee H, Shin J, Hwang SJ, Yang E (2021) Gta: graph truncated attention for retrosynthesis. *Proc AAAI Conf Artif Intell* 35:531–539
46. Sun R, Dai H, Li L, Kearnes S, Dai B (2021) Towards understanding retrosynthesis by energy-based models. *Adv Neural Inf Process Syst* 34:10186–10194
47. Wan Y, Hsieh C-Y, Liao B, Zhang S (2022) Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. In: International Conference on Machine Learning, pp. 22475–22490. PMLR
48. Kim E, Lee D, Kwon Y, Park MS, Choi Y-S (2021) Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *J Chem Inf Model* 61(1):123–133
49. Schneider N, Stief N, Landrum GA (2016) What's what: the (nearly) definitive guide to reaction role assignment. *J Chem Inf Model* 56(12):2336–2346
50. Miyaura N, Suzuki A (1995) Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem Rev* 95(7):2457–2483
51. Paal C (1884) Ueber die derivate des acetophenonacetessigesters und des acetonylacetessigesters. *Ber Dtsch Chem Ges* 17(2):2756–2767
52. Knorr L (1884) Synthese von furfuranderivaten aus dem diacetbernsteinsäureester. *Ber Dtsch Chem Ges* 17(2):2863–2870
53. Borch RF (2003) Reductive amination with sodium cyanoborohydride: N, n-dimethylcyclohexylamine: Cyclohexanamine, 4, 4-dimethyl-. *Org Synth* 52:124–124

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.