Bridging the Gap between Learning and Inference for Diffusion-Based Molecule Generation

Peidong Liu PEIDONG_LIU@STU.SCU.EDU.CN
Wenbo Zhang ZHANGWENBO01@XIDIAN.EDU.CN
Wei Ju JUWEI@SCU.EDU.CN
Jiancheng Lv LVJIANCHENG@SCU.EDU.CN
Xianggen Liu LIUXIANGGEN@SCU.EDU.CN
Sichuan University No. 24 South Section 1, Yihuan Road Chengdu, Sichuan 610065, China

Editors: Hung-yi Lee and Tongliang Liu

Abstract

The paradigm shift toward structure-driven molecule generation has been propelled by advances in deep generative models, such as variational auto-encoders and diffusion models. However, these generative models for molecular design remain constrained by exposure bias, error accumulation, and suboptimal handling of activity cliffs. Here, we introduce DiffGap, a diffusion-based framework that integrates adaptive sampling and pseudo-molecule estimation to bridge the gap between training objectives and inference dynamics in 3D molecule generation. By dynamically aligning intermediate denoising steps with realistic generation trajectories, DiffGap enables the diffusion model to adapt to input biases in advance during the training phase. A temperature annealing module further controls the aligning strength of the adaptive alignment process, ensuring stable learning of the data distribution. Evaluated on the CrossDocked2020 benchmark, DiffGap outperforms existing methods in docking scores and binding affinity, demonstrating superior fidelity in generating drug-like molecules. Our work establishes a principled approach to harmonize generative training with inference mechanics, offering a robust computational toolkit for accelerating structure-based therapeutic discovery. The source code of DiffGap is available at https://github.com/neusymlab/DiffGap.

Keywords: generative model;diffusion model;exposure bias;3D molecule generation

1. Introduction

The pursuit of targeted therapeutic agents represents a cornerstone of modern pharmaceutical research, where molecular design is guided by precise three-dimensional interactions between ligands and disease-associated proteins (Sneader, 2005). This paradigm shift from serendipitous discovery to structure-driven design has been accelerated by advancements in structural biology (Batool et al., 2019; Liu et al., 2021) and computer-aided rational design(Mandal et al., 2009; Jumper et al., 2021; Abramson et al., 2024). In particular, structure-based drug design typically employs molecular docking simulations, pharmacophore modeling, and free energy perturbation calculations to virtual screen promising compounds from curated chemical libraries. While virtual screening (Mayr and Bojanic, 2009; Zhang et al., 2022) remains prevalent, its efficacy is fundamentally constrained by the combinatorial explosion of drug-like chemical space ($\sim 10^{60}$ potential molecules) (Bohacek et al., 1996).

Generative approaches have emerged as promising alternatives to exhaustive screening. Deep generative models (DGMs) attempt to navigate chemical space to generate 2D molecules from learned latent representations (Jin et al., 2018), yet their performance degrades when target properties diverge from training data distributions (Ning et al., 2023). Combinatorial optimization methods (You et al., 2018; Jensen, 2019; Paaßen et al., 2022) show superior potential but remain underexplored for target-specific design. Recent attempts using simulated annealing (Xue et al., 2025) demonstrate limitations in handling activity cliffs—abrupt property changes from minor structural modifications (Pemasinghe and Abeygunawardhana, 2021) due to their myopic optimization strategies.

The integration of geometric deep learning has revolutionized molecular representation learning. Early 2D approaches (Jin et al., 2018; Walters and Barzilay, 2020; Lim et al., 2020) gave way to 3D-equivariant architectures (Fuchs et al., 2020; Satorras et al., 2021) that preserve rotational and translational symmetries critical for molecular geometric structures. Diffusion models leveraging 3D-equivariant networks (Guan et al., 2022, 2023; Huang et al., 2024) to gradually perform the de-noising of the geometric topological structures, achieving state-of-the-art performance in 3D molecule generation. However, the iterative denoising process (typically requiring $\geq$1,000 steps) introduces error accumulation and exposure bias between training objectives and inference conditions analogous to autoregressive sequence generation (Ning et al., 2024).

However, the success of these diffusion models masks a critical vulnerability rooted in their iterative generation process. This vulnerability stems from a fundamental discrepancy between how the models are trained and how they perform inference, and is known as exposure bias. During training, the model always learns to denoise a perfectly conditioned state M_t to a pristine, ground-truth molecule M_0. In contrast, during inference, the model must denoise a state $\hat{M}_t$ which is the result of its own prediction from the previous step, $\hat{M}_{t+1}$. The model is therefore never exposed to its own errors during training, making it fragile when faced with them when autoregressive-style generation (Ning et al., 2024).

To address these limitations, we present DiffGap, a novel diffusion framework incorporating adaptive sampling through pseudo-molecule estimation. Our key insight of the adaptive sampling lies in dynamically aligning the training inputs with the realistic generation trajectories, rather than following teacher-forcing denoising computations. This approach reduces the train-inference discrepancy by treating intermediate predictions as conditional inputs for subsequent steps. We also introduce a temperature annealing module to control the aligning strength of the adaptive sampling process. Extensive evaluations on the CrossDocked2020 benchmark (Francoeur et al., 2020) demonstrate DiffGap's superiority in generating molecules with optimized binding affinities and 3D complementarity.

In summary, our principal contributions are threefold:

- We propose a generative framework, called DiffGap, to address the exposure bias issue of diffusion models by an adaptive sampling strategy.
- The DiffGap framework introduces the construction strategy of pseudo-molecules for training, which is an optimal estimation of the input condition to mimic the inference computations.
- Empirical results show that the molecules generated by DiffGap achieve state-of-the-art docking scores and superior quality on the binding affinity.

2. Related work

2.1. Molecule generation

Existing molecular generation models can be categorized into four groups: string-based, image-based, 2D graph-based, and 3D structure-based. The most common molecular string representation is SMILES (Weininger, 1988), where researchers can reuse language models like RNN and Transformer and quickly apply them to molecular generation tasks following the text approach (Schoenmaker et al., 2023; Brahmavar et al., 2024). For example, researchers trained RNN and its variants on randomized SMILES strings to improve the uniqueness of generated molecules (Grisoni et al., 2020; Arús-Pous et al., 2019) and ChatMol empowered large language model for conversational molecular design diagram (Zeng et al., 2024). 2D molecular image representations (Walters and Barzilay, 2020) and 2D graph representations (Lim et al., 2020; Jin et al., 2018) employed CNNs and GNNs respectively for more atom connection information than string representations. Recently, structure-based models like GraphBP, Pocket2Mol, and diffusion models took the 3D structure and equivariant properties of molecules into account, showing advantages in molecular affinity (Liu et al., 2022; Peng et al., 2022; Guan et al., 2022, 2023; Huang et al., 2024). Our work follows the same denoising theory of diffusion models, but adopts a new training framework (i.e., the adaptive sampling strategy) to improve the generation quality.

2.2. Diffusion models

Introduced by (Sohl-Dickstein et al., 2015) and developed by (Ho et al., 2020; Song et al., 2021), diffusion models have been applied in various fields like unconditional image generation (Ho et al., 2020), text-to-image generation (Nichol et al., 2022). Recently, diffusion models have also been applied to molecular generation tasks, particularly in the field of Structure-Based Drug Design (SBDD). For instance, TargetDiff (Guan et al., 2022) combined with an SE(3)-equivalent network, has surpassed the previous SOTA method, Pocket2Mol (Peng et al., 2022), with a significant docking score on the CrossDock2020 dataset. BindDM adaptively extracted subcomplex and captured the protein-ligand interactions exactly to obtain higher docking affinity than the above models (Huang et al., 2024). The aforementioned models primarily focus on adapting molecular data and properties, proposing various strategies to process such inputs. In contrast, our work centers on the universal sampling strategy of the diffusion models for molecule generation, which is effective for multiple diffusion-based methods, as validated by the empirical results.

2.3. Exposure bias

Exposure bias has been widely studied in sequence generation tasks, particularly in natural language processing (NLP) applications and recommendation algorithms (Yang et al., 2018; Lamb et al., 2016; Zhang et al., 2019). The term exposure bias refers to the discrepancy between training and inference conditions in sequence models. During training, models are conditioned on ground truth data (teacher forcing), while during inference, they generate sequences based on previous predictions, leading to error accumulation (Lamb et al., 2016). Diffusion models share the same problem, but are few solutions (Ning et al., 2023, 2024). To complement research in this domain, we provide a precise prediction of the input condition

to narrow the input gap between training and generation. Different from it, our method provides a more precise prediction of the input condition by Bayesian estimation to narrow the input gap between training and generation.

3. Problem definition

The problem of structure-based 3D molecular generation is defined as a conditional generation process under the specific protein pocket. Formally speaking, a data point consists of pairs of proteins $\mathcal{P}$ and molecular conformations $\mathcal{M}$. The molecular conformation is represented by the concatenation of its atomic 3D Cartesian coordinates $x \in \mathcal{R}^{m\times 3}$ and one-hot encoded atomic types $v \in \mathcal{R}^{m\times k}$ (m denotes the number of atoms in a molecule and k shows the number of potential atomic types), and so is the protein. That is, the goal of structure-based 3D molecular generation is to generate a reasonable molecular conformation $\mathcal{M} = [x, v]$ given the protein $\mathcal{P} = [x_{\mathcal{P}}, v_{\mathcal{P}}]$ using a novel diffusion model with less bias.

4. Methodology

4.1. Classic diffusion process

The diffusion model defines the *diffusion process* with the data $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, which is a Markov chain that incrementally adds Gaussian noise in equation 1 using schedule hyperparameters $\beta_1, \ldots, \beta_T$.

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \qquad q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

Based on the properties of the Gaussian distribution, this process of incrementally adding noise $q(\mathbf{x}_t|\mathbf{x}_0)$ can be simplified in equation 2 with notations $\alpha_t = 1 - \beta_t$; $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

The diffusion model is a parameterized Markov chain that models a latent variable model of the form $p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T})d\mathbf{x}_{1:T}$, used to learn the reverse Gaussian denoising process of the diffusion process $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, i.e., the reverse process. The reverse process can be formalized as a normal distribution in equation 3 because the forward process consists of thousands of steps and each one $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ follows a Gaussian distribution. Consequently, we can utilize $\mu_\theta(\mathbf{x}_t, t)$ and $\Sigma_\theta(\mathbf{x}_t, t)$ to denote the parameters of the normal distribution for a single-step reverse process under the neural network.

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t), \qquad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \tag{3}$$

Intuitively, what we need to know is the denoising distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$ of the data, but it is not tractable. Instead, we can compute the posterior probability of the data point $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ given the original data point as in equation 4.

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}\mathbf{I}) \tag{4}$$

In this way, the ground truth of the denoising probability can be accurately approximated.

4.2. DiffGap framework

DiffGap adopts the diffusion process to learn the conformations M of a molecule in the 3D space, where the molecule state M_t at step t can be determined by the previous state M_{t-1} and the protein structure $\mathcal{P}$.

$$q(M_t|M_{t-1},\mathcal{P}) = P(x_t|x_{t-1},\mathcal{P})P(v_t|v_{t-1},\mathcal{P}) \tag{5}$$

where $M_t = [x_t, v_t]$ is the molecule state at step t. To reduce the computational complexity, we assume the atom coordinates x_t and the identity v_t are independent in the estimation of the transformation probability $q(M_t|M_{t-1},\mathcal{P})$. Similar to equation 1, we model the transformation probability of atomic positions $P(x_t|x_{t-1},\mathcal{P})$ and that of the atomic types $P(v_t|v_{t-1},\mathcal{P})$ using a normal distribution and a categorical distribution (K represents the number of categories), respectively.

The objective of diffusion models is to narrow the divergence between the denoising Gaussian distribution (i.e., ground truth) $q(M_{t-1}|M_t, M_0, \mathcal{P})$ and the predicted distribution $p_\theta(M_{t-1}|M_t,\mathcal{P})$, where the condition M_t is sampled from the unbiased Gaussian distribution.

$$L_{t-1} = \mathbb{E}_q\left[\sum_{t\geq 1} D_{\mathrm{KL}}(q(M_{t-1}|M_t, M_0,\mathcal{P})||p_\theta(M_{t-1}|M_t,\mathcal{P}))\right] \tag{6}$$

$$= \mathbb{E}_q\left[\frac{1}{\Sigma_\theta(M_t,t)}||\tilde{\mu}_t(M_t,M_0) - \mu_\theta(M_t,t)||^2\right] + C \tag{7}$$

where C is a constant and is independent of the neural network.

As for the inference of DiffGap, it starts with Gaussian noises M_T and iteratively denoises from the previous result, finally achieving the goal. That is,

$$p_\theta(\hat{M}_{0:T}|\mathcal{P}) = p(\hat{M}_T)\prod_{t=1}^{T} \overbrace{p_\theta(\hat{M}_{t-1}|\hat{M}_t,\mathcal{P})}^{\text{Condition varies from equation 6.}}, \qquad \hat{M}_t \sim p_\theta(\hat{M}_t|\hat{M}_{t+1},\mathcal{P}) \tag{8}$$

where the condition $\hat{M}_t$ of the denoising process is generated in the last step $p_\theta(M_t|M_{t+1})$, instead of the sample from the unbiased Gaussian distribution.

However, the above inference process of DiffGap will suffer serious exposure bias due to the discrepancy between training and inference. In 3D molecule generation, the conformation space of 3D molecules is huge and rough, with most breaking the chemical and physical rules. A small atom-type shift could result in an unrealistic molecule. During inference, the reverse process can be viewed as an autoregressive generative process iterated multiple times in time steps, generating M_0 by gradually denoising M_t sampled randomly according to equation 3. However, general diffusion models require a large number of iteration steps (typically 1000), leading to error accumulation and exposure bias issues.

Adaptive sampling strategy. To address the exposure bias issue in traditional diffusion models, we propose an adaptive sampling strategy, which reduces the discrepancy in data distribution between training and inference by introducing reasonable noise into

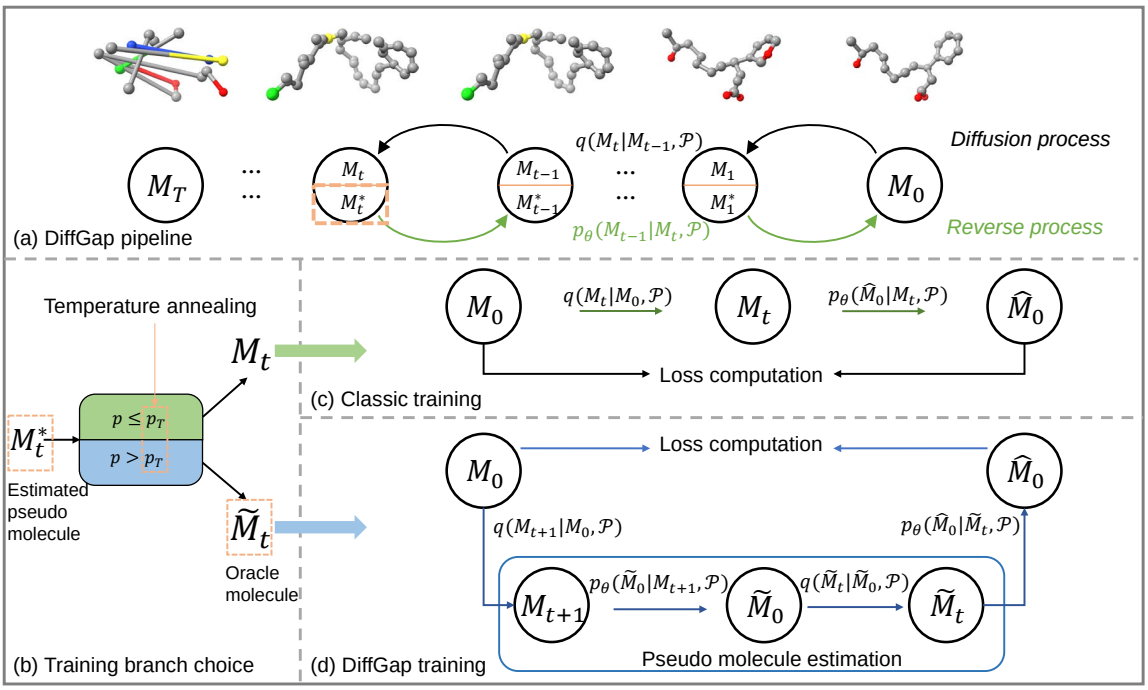

Figure 1: The overview of DiffGap pipeline under target-aware molecule generation task. DiffGap is consistent with the logic of DDPM, with the difference lying in the reverse process during training, shown in Figure (a). In the reverse process of training, the ground truth is selected probabilistically between the original ground truth (denoted as M_i like M_t) and the model's real-time predicted value (denoted as $\widetilde{M}_i$ like $\widetilde{M}_t$), with a probability p_T favoring the original value in Figure (b). What's more, the probability p_T is periodically updated by temperature annealing. Figures (c) and (d) display the two training pathways guided by p_T and we will discuss it further in section 4.

the training phase. The core idea of DiffGap is to utilize model-predicted conformations as ground truth during training probabilistically. In particular, we denote M_t^* as a perturbed condition in the denoising probability $q(M_{t-1}|M_t, M_0)$ to replace the ground truth sample M_t, which is unavailable in the inference phase. Therefore, the objective function of DiffGap is

$$L_{t-1} = \mathbb{E}_q \left[\sum_{t \geq 1} D_{\mathrm{KL}}(q(M_{t-1}|M_t^*, M_0, \mathcal{P}) || p_\theta(M_{t-1}|M_t^*, \mathcal{P})) \right], \quad M_t^* = \mathrm{Perturbation}(M_t) \tag{9}$$

Since molecule $M_t = [x_t, v_t]$ has two components, we need to compute the atom coordinate loss and the atom type loss respectively using equation 6. We measure the distance between the predicted atom coordinates and the ground truth for coordinate loss, and take the KL-divergence of categorical distributions (i.e., $c(v_t, v_0)$) for the atom type loss. Then

the objective function can be further transformed to

$$L_{t-1} = L_{t-1}^{(x)} + \lambda L_{t-1}^{(v)}, \qquad L_{t-1}^{(x)} = \gamma_t ||x_0 - \hat{x}_0||^2 + C, \qquad L_{t-1}^{(v)} = \sum_k c(v_t, v_0)_k \log \frac{c(v_t, v_0)_k}{c(v_t, \hat{v}_0)_k} \tag{10}$$

where x_0 and v_0 are the atom coordinate and atom type of M_0, respectively[1]. $\hat{x}_0$ and $\hat{v}_0$ are the predictions of the diffusion model ϕ, given by $\gamma_t = \frac{\bar{\alpha}_{t-1}\beta_t^2}{2\Sigma_\theta(x_t,t)(1-\bar{\alpha}_t)^2}$ and C is a constant. Note that, we adopt $\phi_\theta(M_t^*, t, \mathcal{P})$ as the diffusion model and the denoising probability $p_\theta(M_{t-1}|M_t^*, \mathcal{P}))$ can be derived based on it.

$$\hat{M}_0 = [\hat{x}_0, \hat{v}_0] = \phi_\theta(M_t^*, t, \mathcal{P}) \tag{11}$$

Pseudo molecule estimation. The main focus of this paper is to figure out an approximated sample representation M_t^* that can bridge the gap between the training and inference of diffusion-based molecule generation. Therefore, M_t^* should satisfy the following two properties. On the one hand, M_t^* is expected to contain the key information of estimated pseudo molecule at the t-th step (i.e., M_t). On the other hand, M_t^* must resemble the iterative output of the diffusion model.

To this end, we propose pseudo molecule estimation, whose core idea is to regard the model's current predictions as the ground truth instead of using the standard noisy sample in training time. Using this estimate allows the model to make the data distribution during training a weighted average over the true distribution and the model's learned distribution, thus reducing the gap between training and inference.

Concretely, the pseudo molecule estimation leverages the diffusion model to re-predict the original molecule state based on the molecule state at step t. As a result, an estimated original molecule state is obtained.

$$\widetilde{M}_0 = \phi_\theta(M_t, t, \mathcal{P}) \tag{12}$$

Next, we apply Bayesian theory and add noises into the estimated original molecule $\widetilde{M}_0$, thus yielding a new variable $\widetilde{M}_t$ that describes the molecule state at step t.

$$\widetilde{M}_t \sim q(\widetilde{M}_t|\widetilde{M}_0, \mathcal{P}) \tag{13}$$

In this way, the estimated original molecule state $\widetilde{M}_0$ mimics the scenario in the inference stage where the true molecule state is not available. And $\widetilde{M}_t$ is a pseudo molecule state at step t, which can satisfy the two properties of M_t^*. However, we do not directly use the estimated molecule state $\widetilde{M}_0$ as M_t^*, because purely adopting the model prediction as the condition would contaminate the training process. Instead, M_t^* picks M_t with probability p_T, otherwise chooses the estimated molecule $\widetilde{M}_t$. Hence, the training process of the diffusion model can be controlled by the pre-defined probability p_T.

Note that although the pseudo molecule estimation applies the diffusion model ϕ to produce the condition M_t^*, this process does not need to receive gradients and optimize.

1. $\mathbf{x}_0$ and x_0 are different here. The former represents a general data sample in the classic diffusion process while the latter stands for the atom coordinate of the original molecule M_0

That is, the pseudo molecule estimation will dynamically use the diffusion model ϕ to add noises into the condition in the training stage and thus narrow the difference between the learning and inference.

$$P(M_t^*) = \begin{cases} p_T, & M_t^* = M_t \\ 1 - p_T, & M_t^* = \widetilde{M}_t \end{cases} \tag{14}$$

Probability temperature annealing. Hoping the model can learn the data distribution more smoothly during training, we use monotonic decline functions to control the selection probability. Intuitively, we need this probability curve to have a lower cooling rate at the beginning of training, that is, a smaller derivative value, and a higher cooling rate at the end of training, so that it can quickly adapt to changes from training to inference. We take the OR (Zhang et al., 2019) model's curve as the first one, which goes with a hyperparameter (equation 15) and is borrowed from Bengio (Bengio et al., 2015). However, under its default setting (shown in Setup), the probability p_T cools too quickly, which is not conducive to learning a robust data distribution.

$$p_T = \frac{\mu}{\mu + \exp(e/\mu)} \tag{15}$$

To make the learning process more stable, we propose two other temperature anneals, one is linear annealing (equation 16 left) and the other is arc annealing (equation 16 right). For the sake of conciseness, this can also be regarded as $p = \sqrt{r^2 - e^2}/r)$, we use a modified quarter circle curve as the cooling curve, make the result a real number and finally regularize it to ensure that the curve value range is in $[0, 1]$). In order to avoid p_T being too low in the later stages of training, the model basically enters the self-verification learning stage, which has been reinforcing bias and making it difficult to learn the original distribution. We use $\min(p, \text{lower\_bound})$ to obtain the probability of actual use.

$$p_T = 1 + \text{slope} * e, \qquad p_T = \sqrt{\max(r^2 - (e/100)^2, 0)}/r \tag{16}$$

5. Experiments

5.1. Setup

Data. We use CrossDocked2020 (Francoeur et al., 2020), the commonly used protein-ligand pairs dataset, as a benchmark dataset for both training and evaluation. Similar to (Luo et al., 2021; Guan et al., 2022, 2023; Huang et al., 2024), we filter the complexes with RMSD (Root Mean Square Deviation, the measure of the average distance between the atoms of superimposed molecules) higher than 1 Å and remains 100,000 pairs for training, 100 pairs for testing.

Baseline. We use our model to compare the affinity of the generated molecules with liGAN (Ragoza et al., 2022), AR (Luo et al., 2021), Pocket2Mol (Peng et al., 2022), GraphBP (Liu et al., 2022), TargetDiff (Guan et al., 2022), DecompDiff (Guan et al., 2023) and BindDM (Huang et al., 2024). In particular, TargetDiff, DecompDiff, and BindDM represent previous state-of-the-art performance in 3D molecule generation to a given protein structure with diffusion process and equivariant graph neural network, considering rotational and translational equivariance.

Table 1: We compared our models with +Ours sequentially with all other models, indicating the best performing method in each case in **bold** and highlighting the metrics where our method achieved second place with underlining. In the subsequent text, we will use DiffGap to represent BindDM+Ours.

	Vina Score(↓)		Vina Min(↓)		Vina Dock(↓)		High Affinity(↑)		QED(↑)		SA(↑)		Div(↑)	
Models	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.	Avg.	Med.
Reference	-6.36	-6.46	6.71	-6.49	-7.45	-7.26	-	-	0.48	0.47	0.73	0.74	-	-
liGAN	-	-	-	-	-6.33	-6.20	21.1%	11.1%	0.39	0.39	0.59	0.57	0.66	0.67
GraphBP	-	-	-	-	-4.80	-4.70	14.2%	6.7%	0.43	0.45	0.49	0.48	**0.79**	**0.78**
AR	-5.75	-5.64	-6.18	-5.88	-6.75	-6.62	37.9%	31.0%	0.51	0.50	0.63	0.63	0.70	0.70
Pocket2Mol	-5.14	-4.70	-6.42	-5.82	-7.15	-6.79	48.4%	51.1%	**0.56**	**0.57**	**0.74**	**0.75**	0.69	0.71
DecompDiff	-5.67	-6.04	-7.04	-7.09	-8.39	-8.43	64.4%	71.0%	0.45	0.43	0.61	0.60	0.68	0.68
TargetDiff	-5.47	-6.30	-6.64	-6.83	-7.80	-7.91	58.1%	59.1%	0.48	0.48	0.58	0.58	0.72	0.71
+Ours	**-6.51**	**-7.18**	**-7.50**	**-7.38**	**-8.54**	**-8.38**	59.0%	62.8%	0.46	0.46	0.56	0.56	**0.79**	0.77
BindDM	-5.92	-6.81	-7.29	-7.34	-8.41	-8.37	64.8%	71.6%	0.51	0.52	0.58	0.58	0.75	0.74
+Ours (DiffGap)	**-6.28**	**-6.90**	**-7.39**	**-7.45**	**-8.43**	**-8.47**	**68.9%**	**72.2%**	0.51	0.52	0.59	0.58	0.75	0.75

DiffGap. We employ the E(n)-Equivariant GNN (Satorras et al., 2021) as the backbone model, which contains 9 equivariant layers and makes diffusion steps $T = 1000$ the same as DDPM. As liGAN and AR do, we choose OpenBabel (O'Boyle et al., 2011) to reconstruct the 3D molecule from atom coordinates. We use the Adam optimizer, with $\beta = (0.95, 0.999)$ and a learning rate of 1e-4, without weight decay. We set the max training step to 200K and the batch size to 4 for all the models. In the scenario, the probability of not computing the prediction $P(t = T - 1)^{\text{batchsize}}$ is 0.996, which is acceptable and can be ignored. More details can be found in the README file within the code repository.

5.2. Results

We conduct several experiments to compare our method against the aforementioned baseline models, primarily evaluating the performance of our approach in terms of the quality of generated molecules. The main data are presented in four aspects: chemical bond distributions, binding affinity, and molecular properties of DiffGap. Additionally, we perform ablation experiments to verify the rationale behind the hyperparameter selection of the adaptive sampling strategy.

Bond distribution. First, we consider the distribution of atomic coordinates and the most common types of bonds connected to carbon atoms after applying the adaptive sampling strategy. In particular, the bond types include carbon-carbon bonds, carbon-nitrogen bonds, and carbon-oxygen bonds, covering single bonds('-'), double bonds('='), and aromatic bonds(':'). We evaluate the difference between the generated bonds and the reference bonds using the Jensen-Shannon divergence (Lin, 1991), where a lower value indicates better performance.

As shown in Table 2, the molecules generated by our method are among the best three results, comparable to BindDM and surpassing other models. Meanwhile, our approach exhibits a significant advantage across all carbon bonds, particularly in double and aromatic bonds, where the latter achieves a twofold improvement over the previous best result. We see in Figure 2 that DiffGap achieves the lowest JSD of 0.065 to reference in all-atom pairs

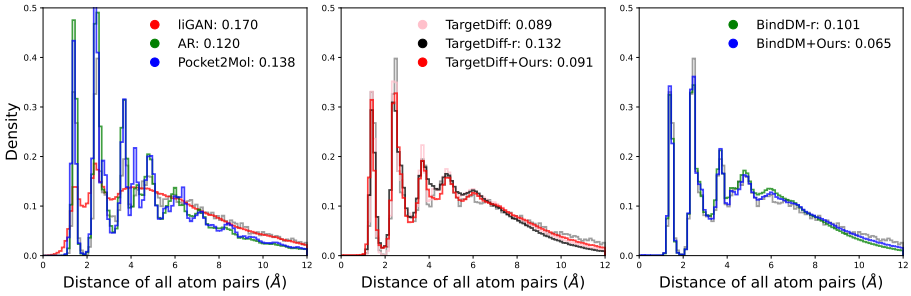

Figure 2: Comparing the distribution for distances of all-atom for reference molecules in the test set (gray) and generated molecules (color). Jensen-Shannon divergence (JSD↓) between two distributions is reported.

distance distribution of the ligands in test sets, exceeding notably TargetDiff and others. Moreover, TargetDiff+Ours beats TargetDiff-r (the reproduced TargetDiff). That is to say, the molecules generated by DiffGap surpass those produced by prior methods overall.

Table 2: Jensen-Shannon divergence between the bond distance for reference and the generated. And we highlight the best 3 results with **bold text**, underlined text, and *italic text* respectively.

Bond(↓)	liGAN	AR	Pocket2Mol	TargetDiff	DecompDiff	BindDM	DiffGap
C-C	0.601	0.609	0.496	*0.369*	0.359	0.380	**0.357**
C-N	0.634	0.474	0.416	0.363	*0.344*	0.265	**0.253**
C-O	0.656	0.492	0.454	0.421	0.376	**0.329**	*0.388*
C=C	0.665	0.620	0.561	*0.505*	0.537	0.229	**0.189**
C=N	0.749	0.635	0.629	*0.550*	0.584	**0.245**	0.260
C=O	0.661	0.558	0.516	0.461	0.374	**0.249**	*0.376*
C:C	0.497	0.451	0.416	*0.263*	0.251	0.282	**0.141**
C:N	0.638	0.552	0.487	0.235	0.269	**0.130**	*0.240*

Binding affinity. The binding affinity between a molecule and a protein is measured by the energy released after binding. AutoDock Vina (Eberhardt et al., 2021) is usually used to calculate the energy, which acts as a crucial evaluation metric. Therefore, we use the docking scores and the results compared with the reference as metrics to compare and evaluate all the models,: Vina Score, Vina Min, Vina Dock, and High Affinity. Vina Score is used to evaluate the stability of the small molecule-protein binding, Vina Min represents the minimum value during the docking process, Vina Dock energy attempts to find the lowest energy binding conformation, and High Affinity exhibits the proportion of superiority over the reference on Vina Dock. As indicated in Table 1, DiffGap obtains higher mean and median in all affinity-related metrics compared to the baselines with the highest improvement reaching 4.1% in the mean of High Affinity. Otherwise, we reorder the median Vina Dock in ascending order according to our method for eight methods. As evidenced by Figure 3, DiffGap enables substantially superior performance compared to

existing methods, yielding a minimum 20% enhancement over TargetDiff and exceeding BindDM by more than 30%.

Molecular properties. As for property-related metrics, drug-likeness QED (Bickerton et al., 2012), synthesizability SA (Ertl and Schuffenhauer, 2009) and diversity are commonly used for evaluation. Unlike that DecompDiff accepts a trade-off between property-related metrics and affinity-related metrics, we maintain proper properties and make somewhat progress in the mean of SA and the median of diversity in Table 1. Nevertheless, we put less attention on QED and SA because many invalid molecules will be filtered out by virtual screening and it would be acceptable in a reasonable range.

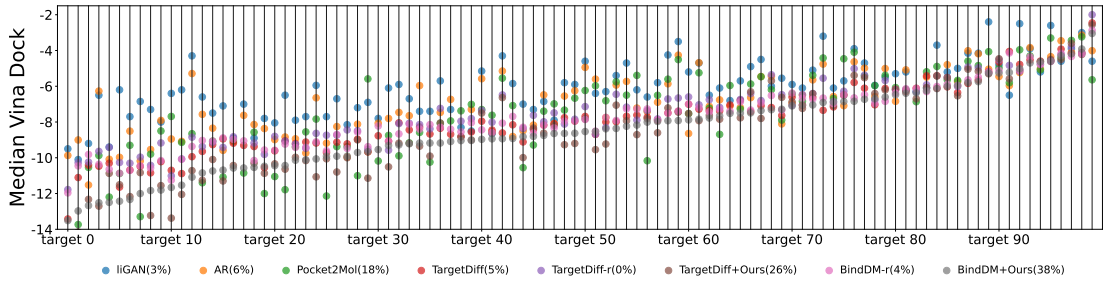

Figure 3: Median Vina dock energy for five models across 100 testing targets. The percentage represents the proportion of the model achieving the best binding affinity.

5.3. Ablation study

These are two ablation studies to validate the rationality of our method. The first study focuses on the selection of annealing strategies, while the second involves the specific parameter settings of the selected annealing method.

Choice of annealing strategy. We employ ablation experiments for comparison to filter out the best method, which considers the impact of both the lower bound and the annealing method. Due to the time-consuming nature of sampling and evaluating the full test dataset of 100 targets, we randomly selected 10 targets for the ablation experiments. Then we used the docking scores as the evaluation metrics and calculated the time cost multiplier of DiffGap compared to the classic way (marked as Cost in Table 3).

Overall, arc annealing outperformed the others at both lower bounds of 0.5 and 0.8, with it performing best when the lower bound was set to 0.5 in Table 3. Consequently, we ultimately choose arc annealing and set the lower bound to 0.5, and we think the extra time cost is acceptable.

Arc annealing comparison. We initially hypothesized that the arc annealing curve would yield the best results when $r = 2$, although it remains to be validated. Therefore, we design a series of ablation experiments, selecting the values 1.5, 2, 3, 4, 8, and infinity to empirically determine the optimal value for the hyperparameter r. The experimental outcomes confirm our initial hypothesis.

When r equals infinity, p_T is 1, which corresponds to not using our method at all, resulting in the lowest Vina scores, thereby proving the effectiveness of our method. When

Table 3: Annealing methods comparison across 10 testing targets.

Metrics / Annealing	Lower Bound	Vina Score(↓) Avg.	Vina Score(↓) Med.	Vina Min(↓) Avg.	Vina Min(↓) Med.	Vina Dock(↓) Avg.	Vina Dock(↓) Med.	Cost(↓) Avg.	Cost(↓) Avg.
Original (equation 15)	0.5	-4.588	-4.705	-4.898	-4.893	-5.661	-5.540	-5.047	1.38
	0.8	-4.209	-4.343	-4.544	-4.580	-5.471	-5.392	-4.757	1.18
Linear (equation 16 left)	0.5	-6.564	**-6.499**	-6.617	-6.241	-7.421	-7.249	-6.765	1.37
	0.8	**-6.662**	-6.418	-6.593	-6.223	-7.317	-7.146	-6.727	1.18
Arc (equation 16 right)	0.5	-6.496	-6.384	**-6.780**	**-6.525**	**-7.470**	**-7.369**	**-6.837**	1.19
	0.8	-6.413	-6.189	-6.719	-6.445	-7.463	-7.398	-6.771	**1.12**

r is set to 1.5, 3, 4, or 8, the results are comparable, indicating that even with the early introduction of estimation or fewer opportunities for estimation, there is still a certain degree of improvement. When r equals 2, the best results are achieved, surpassing TargetDiff.

Table 4: Annealing comparison for better hyper-parameter.

Metrics / r	Vina Score(↓) Avg.	Vina Score(↓) Med.	Vina Min(↓) Avg.	Vina Min(↓) Med.	Vina Dock(↓) Avg.	Vina Dock(↓) Med.	Cost(↓) Avg.
∞	-5.41	-6.01	-6.32	-6.28	-7.28	-7.39	1
8	-5.55	-6.27	-6.49	-6.54	-7.35	-7.51	1.01
4	-5.58	-6.10	-6.49	-6.54	-7.35	-7.51	1.04
3	-5.61	-6.27	-6.49	-6.48	-7.53	-7.53	1.08
2	**-6.51**	**-7.18**	**-7.50**	**-7.38**	**-8.54**	**-8.38**	1.12
1.5	-5.63	-6.12	-6.41	-6.37	-7.32	-7.42	1.14

6. Conclusion

DiffGap addresses critical limitations in diffusion-based molecular generation through adaptive sampling and pseudo-molecule estimation. By dynamically aligning training trajectories with inference dynamics to mitigate exposure bias and error accumulation, the framework ensures stable learning of 3D molecular distributions while preemptively adapting to input biases.Evaluations on CrossDocked2020 confirm DiffGap's superiority in generating molecules with enhanced docking scores and binding affinities, outperforming existing methods. The primary impact of this work is its versatility as a general-purpose framework that can be integrated as a plug-in to enhance various diffusion-based architectures. By generating molecules with superior docking scores and binding affinities, DiffGap offers a robust computational toolkit to accelerate structure-driven drug discovery. However, we acknowledge that this mechanism is specific to the diffusion model and that the underlying principles are not yet understood, making it difficult to continue optimizing.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62206192), the Natural Science Foundation of Sichuan Province (No. 2023NSFSC1408, 2024NSFTD0048),

and the Science and Technology Major Project of Sichuan Province (No. 2024ZDZX0003). We also acknowledge the support of Sichuan Province Engineering Technology Research Center of Broadband Electronics Intelligent Manufacturing.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

Josep Arús-Pous, Simon Viet Johansson, Oleksii Prykhodko, Esben Jannik Bjerrum, Christian Tyrchan, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Randomized smiles strings improve the quality of molecular generative models. *Journal of cheminformatics*, 11:1–13, 2019.

Maria Batool, Bilal Ahmad, and Sangdun Choi. A structure-based drug discovery paradigm. *International journal of molecular sciences*, 20(11):2783, 2019.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.

Regine S Bohacek, Colin McMartin, and Wayne C Guida. The art and practice of structure-based drug design: a molecular modeling perspective. *Medicinal research reviews*, 16(1): 3–50, 1996.

Shreyas Bhat Brahmavar, Ashwin Srinivasan, Tirtharaj Dash, Sowmya Ramaswamy Krishnan, Lovekesh Vig, Arijit Roy, and Raviprasad Aduri. Generating novel leads for drug discovery using llms with logical feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21–29, 2024.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

Paul G Francoeur, Tomohide Masuda, Jocelyn Sunseri, Andrew Jia, Richard B Iovanisci, Ian Snyder, and David R Koes. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of chemical information and modeling*, 60(9):4200–4215, 2020.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020.

Francesca Grisoni, Michael Moret, Robin Lingwood, and Gisbert Schneider. Bidirectional molecule generation with recurrent neural networks. *Journal of chemical information and modeling*, 60(3):1175–1183, 2020.

Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

Jiaqi Guan, Xiangxin Zhou, Yuwei Yang, Yu Bao, Jian Peng, Jianzhu Ma, Qiang Liu, Liang Wang, and Quanquan Gu. Decompdiff: Diffusion models with decomposed priors for structure-based drug design. In *International Conference on Machine Learning*, pages 11827–11846. PMLR, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Zhilin Huang, Ling Yang, Zaixi Zhang, Xiangxin Zhou, Yu Bao, Xiawu Zheng, Yuwei Yang, Yu Wang, and Wenming Yang. Binding-adaptive diffusion models for structure-based drug design. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12671–12679, 2024.

Jan H Jensen. A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space. *Chemical science*, 10(12):3567–3572, 2019.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

Jaechang Lim, Sang-Yeon Hwang, Seokhyun Moon, Seungsu Kim, and Woo Youn Kim. Scaffold-based molecular design with a graph generative model. *Chemical science*, 11(4): 1153–1164, 2020.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. In *International Conference on Machine Learning*, pages 13912–13924. PMLR, 2022.

Xianggen Liu, Yunan Luo, Pengyong Li, Sen Song, and Jian Peng. Deep geometric representations for modeling effects of mutations on protein-protein binding affinity. *PLoS computational biology*, 17(8):e1009284, 2021.

Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.

Soma Mandal, Sanat K Mandal, et al. Rational drug design. *European journal of pharmacology*, 625(1-3):90–100, 2009.

Lorenz M Mayr and Dejan Bojanic. Novel trends in high-throughput screening. *Current opinion in pharmacology*, 9(5):580–588, 2009.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *International Conference on Machine Learning*, pages 26245–26265. PMLR, 2023.

Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=xEJMoj1SpX.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3:1–14, 2011.

Benjamin Paaßen, Irena Koprinska, and Kalina Yacef. Recursive tree grammar autoencoders. *Machine Learning*, 111(9):3393–3423, 2022.

Sajeewa Pemasinghe and Pradeep KW Abeygunawardhana. Simulated annealing and it's application in molecular structure optimizations. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 258–262. IEEE, 2021.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pages 17644–17655. PMLR, 2022.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13 (9):2701–2713, 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pages 9323–9332. PMLR, 2021.

Linde Schoenmaker, Olivier JM Béquignon, Willem Jespers, and Gerard JP van Westen. Uncorrupt smiles: a novel approach to de novo design. *Journal of Cheminformatics*, 15 (1):22, 2023.

Walter Sneader. *Drug discovery: a history*. John Wiley & Sons, 2005.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

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.

Zhe Xue, Chenwei Sun, Wenhao Zheng, Jiancheng Lv, and Xianggen Liu. TargetSA: adaptive simulated annealing for target-specific drug design. *Bioinformatics*, 41(1):btae730, 2025.

Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems*, pages 279–287, 2018.

Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *Advances in neural information processing systems*, 31, 2018.

Zheni Zeng, Bangchen Yin, Shipeng Wang, Jiarui Liu, Cheng Yang, Haishen Yao, Xingzhi Sun, Maosong Sun, Guotong Xie, and Zhiyuan Liu. Chatmol: interactive molecular discovery with natural language. *Bioinformatics*, 40(9):btae534, 2024.

Baohua Zhang, Hui Li, Kunqian Yu, and Zhong Jin. Molecular docking-based computational platform for high-throughput virtual screening. *CCF Transactions on High Performance Computing*, pages 1–12, 2022.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, July 2019.