Rethinking Specificity in SBDD: Leveraging Delta Score and Energy-Guided Diffusion

Bowen Gao [*1] **Minsi Ren** [*2 3] **Yuyan Ni** [4] **Yanwen Huang** [5] **Bo Qiang** [5]
Zhi-Ming Ma [4] **Wei-Ying Ma** [1] **Yanyan Lan** [1 6 7]

Abstract

In the field of Structure-based Drug Design (SBDD), deep learning-based generative models have achieved outstanding performance in terms of docking score. However, further study shows that the existing molecular generative methods and docking scores both have lacked consideration in terms of specificity, which means that generated molecules bind to almost every protein pocket with high affinity. To address this, we introduce the *Delta Score*, a new metric for evaluating the specificity of molecular binding. To further incorporate this insight for generation, we develop an innovative energy-guided approach using contrastive learning, with active compounds as decoys, to direct generative models toward creating molecules with high specificity. Our empirical results show that this method not only enhances the delta score but also maintains or improves traditional docking scores, successfully bridging the gap between SBDD and real-world needs.

1. Introduction

Structure-based drug design (SBDD) has become a pivotal strategy in creating novel therapeutic agents. This approach leverages the three-dimensional structural information of target receptors to generate drug-like small molecules that can bind to these receptors effectively. Recent advancements, particularly those in deep learning-based generative models, have shown promising results. Studies such as (Luo et al., 2021; Long et al., 2022; Guan et al., 2023a;b; Peng et al., 2022) demonstrate the capability of these models to achieve docking scores that surpass those of reference ligands.

However, a deeper analysis suggests that these accomplishments may not fully reflect practical efficacy. Further investigation of generated molecules with high docking scores reveals a critical issue: not only do they have a high docking score with the assigned pocket, but they also achieve high docking scores with other unintended protein pockets. From a biological perspective, this observation indicates that current deep learning generative models generate molecules of low specificity.

Specificity is a crucial aspect of drug design. The efficacy of a drug is not solely dependent on its ability to bind to the intended disease-related target but also on its specificity for not binding to essential proteins. Promiscuous drugs, interacting with multiple biological targets, can potentially cause serious adverse effects(Harrison, 2016; Lin et al., 2019; Wong et al., 2019). This lack of specificity contributes to the development of pan-assay interference compounds (PAINS), characterized by their undesirable broad biological activity and potential toxicity (Schneider & Schneider, 2016).

Unfortunately, specificity has been a largely overlooked factor in benchmarking SBDD previously. Currently, the most important metric for SBDD is based on docking scores. Traditional scoring functions merely evaluate the docking pose's spatial and electrostatic fit, failing to differentiate between highly specific molecules and promiscuous ones that bind to multiple targets (Zheng et al., 2022). This limitation means modifications can artificially boost docking scores, misleadingly suggesting a compound's specificity and effectiveness. Thus, there's a pressing need for new metrics that accurately assess the specificity of molecule-pocket interactions, moving beyond the generalized interaction probability that current docking scores offer.

Besides benchmarking, they also fail to explicitly model the specific binding behavior for the generation. Generally,

[*]Equal contribution [1]Institute for AI Industry Research (AIR), Tsinghua University. [2]Institute of Automation, Chinese Academy of Sciences. [3]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. [4]Academy of Mathematics and Systems Science, Chinese Academy of Sciences. [5]Department of Pharmaceutical Science, Peking University. [6]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China. [7]Beijing Academy of Artificial Intelligence (BAAI) . Correspondence to: Yanyan Lan <lanyanyan@air.tsinghua.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

during training, these methods treat the pocket as context and attempt to reconstruct the ligand. This approach emphasizes maximizing the joint probability of pocket-molecule pairs rather than the conditional probability of a molecule given a pocket, which more precisely reflects a drug's specificity and efficacy. We denote the conditional probability of a molecule given a pocket as the conditional binding probability.

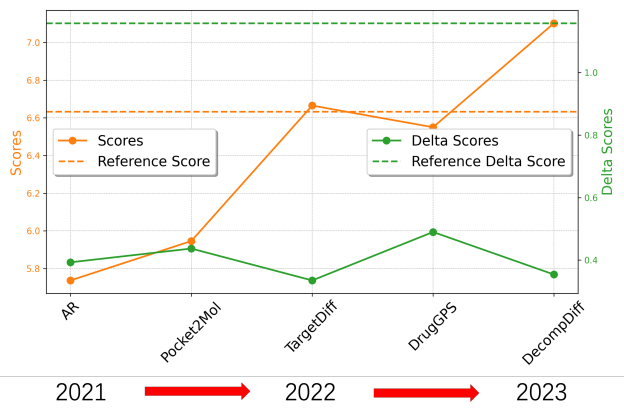


Figure 1. The evolution of absolute docking scores and delta scores obtained by various methods, organized chronologically.

To address these shortcomings, we have adopted a probabilistic framework and defined the conditional binding probability. From this perspective, we introduce a better evaluation metric named *Delta Score* to measure the specific binding ability of current deep-learning-based SBDD models. The Delta Score has been designed and theoretically substantiated to reflect a molecule's selective binding behavior more accurately. It effectively bridges the theoretical predictions with the practical necessities of drug design.

Our analysis using the Delta Score has led to an essential finding: despite their prowess in achieving high docking scores, deep learning models underperform compared to reference ligands when evaluated using the Delta Score, as shown in Figure 1. This discrepancy highlights the need for improved deep-learning-based drug design methods.

In response to these challenges, we have developed an energy-guided approach(Dhariwal & Nichol, 2021; Zhao et al., 2022; Bao et al., 2022). This approach is anchored in the use of contrastive learning and the strategic use of active compounds as decoys during training, inspired by (Radford et al., 2021; Gao et al., 2023). The training objective of our energy function aligns with the conditional binding probability formula, meaning that minimizing the training loss is equivalent to maximizing the conditional binding probability. This function guides the diffusion process in our generative model, directing it toward the conditional generation of molecular structures that exhibit high specific binding behavior.

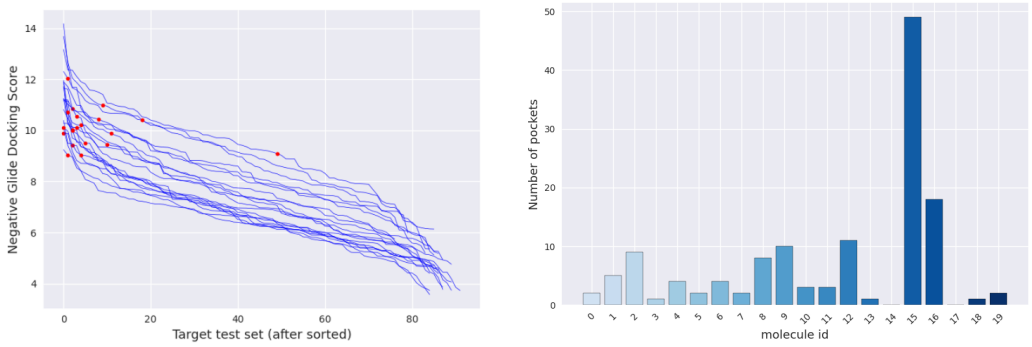
Our experimental results are encouraging, demonstrating that our method not only improves the Delta Score but also maintains or enhances conventional docking scores. By implementing this novel approach, we aim to refine the process of AI-assisted drug design. Our goal is to ensure that the resulting molecules are not only effective in binding to their targets but also exhibit the selectivity required for safe and efficacious therapeutic use.

Our contributions are summarized as: 1) The proposal of the Delta Score, an evaluation metric, measures the specific binding ability of generated molecules, demonstrating that advancements in previous methods predominantly enhance unconditional aspects rather than conditional binding. 2) The introduction of a specific binding energy guidance approach, which explicitly models specific binding, directs the diffusion process toward producing molecules with a higher conditional binding probability to their designated targets. 3) The achievement of promising empirical outcomes, which not only improve the Delta Score but also enhance the traditional docking score.

2. Related Works

With the emergence of geometric deep learning models (Satorras et al., 2021; Hoogeboom et al., 2022; Geiger & Smidt, 2022), the field of Structure-Based Drug Design (SBDD) has shifted towards 3D neural networks for encoding protein structures and decoding 3D molecule conformations, representing real-world 3D interactions. Various methods have been proposed, including voxel-based methods (Masuda et al., 2020), auto-regressive models (Luo et al., 2021; Long et al., 2022), and diffusion models (Guan et al., 2023a;b).

As for the evaluation metrics, most of the previous work used Vina score (Trott & Olson, 2010) for binding affinity evaluation, which is a typical docking software to predict the interactions between a small molecule and a protein target. However, several studies have shown that other commercial docking software like Glide (Friesner et al., 2004) and Gold (Verdonk et al., 2003), achieves a more accurate docking pose predicton and virtual screening ability compared to Vina. Recently, several studies have pointed out the issues of current benchmarking baselines, like PoseCheck(Harris et al., 2023) and PoseBusters (Buttenschoen et al., 2024). They mainly focus on the reliability of the conformations directly generated by the generative models. However, there is limited discussion on the reliability and value of the widely used docking score metric in the assessment of SBDD methods.



(a) Docking Scores against targets in the test set after sorted (multiplied by -1)

(b) The number of pockets which surpass true target on docking score

Figure 2. In the left image, each line represents the sorted docking score of a small molecule against all pockets in the testset. The highlighted red represents its true target. The right image displays the number of test pockets for each small molecule in which the docking score is higher than their respective true target.

3. Reassessing SBDD: Addressing Molecular Specificity

We first conduct experiments on CrossDocked2020 dataset (Francoeur et al., 2020) to compare different deep generative models, including auto-regressive model (denoted as AR) (Luo et al., 2021), Pocket2mol (Peng et al., 2022), DrugGPS (Zhang & Liu, 2023), Targetdiff (Guan et al., 2023a) and Decompdiff (Guan et al., 2023b). The data preprocessing and splitting are all following Guan et al. (2023a) to ensure a fair comparison. For each method, we generate 20 molecules for each target pocket in the testset. Besides, we randomly sample 100 molecules for each target from the trainset, to form a baseline named 'Random'. We adopt Glide which has demonstrated superior performance in predicting both accurate conformation and binding affinity to make our results more convincing (Zhang et al., 2023).

3.1. Notable Findings in Docking Scores

The analysis of Table 1 reveals several key observations: 1. All methods can generate molecules with docking scores surpassing those of reference ligands across different targets. 2. Diffusion based methods consistently exceed the reference ligands in mean docking scores. 3. Surprisingly, **random sampling** from the training set can yield a mean Vina score close to that of reference ligands, with 30.83% of molecules performing better.

Table 1 shows that modern deep generative models are highly effective in producing molecules that outperform reference ligands, often with a success rate of 50% or more. This suggests significant progress in Structure-Based Drug Design (SBDD). However, the reliability of docking scores as a sole quality measure is questionable. Research points to the susceptibility of docking software to manipulation, and our findings support this — about 30% of randomly selected molecules from a molecular library can outperform reference ligands in docking scores, raising concerns about the current docking score metrics' validity.

Table 1. Docking Score Experimental Results. For a pocket with multiple generated molecules, we show results with both mean and min aggregation values.

	Docking Scores(↓)		better than Ref (↑)
Methods	mean	min	
Random	-5.727	-8.886	30.83%
AR	-5.738	-7.529	30.55%
Pocket2Mol	-5.945	-7.734	32.48%
DrugGPS	-6.554	-8.732	48.13%
TargetDiff	-6.664	-8.501	50.06%
DecompDiff	-7.102	-8.970	56.04%
Reference ligand	-6.632	-6.632	-

3.2. Specificity Concerns in SBDD

Our examination of molecules **with superior scores** from current models reveals a significant issue with specificity. We generated 100 molecules for each of 20 randomly selected pockets in the test set and chose the one with the best docking score for each pocket, resulting in 20 molecules. Despite their high glide docking scores (over -9) on their intended targets, cross-docking these molecules with all test set pockets (figure 2) showed that only 2 out of 20 had the highest docking score for their true targets. This indicates that these molecules generally bind well to multiple pockets, not just their specific target, highlighting a lack of specificity.

Specificity is crucial in drug design; non-specific molecules can bind to unintended targets, potentially causing adverse effects and diminishing efficacy. This issue is compounded by pseudo-active sites (PAINS), which can falsely enhance docking scores and mislead predictions of a molecule's

true target. The prevalence of non-specific binding not only compromises drug effectiveness but also raises safety concerns due to potential side effects. High specificity in drug development is essential to ensure selective binding, maximizing therapeutic benefit and minimizing risks.Both previous models and evaluation metrics have overlooked this critical aspect of specificity.

3.2.1. Inadequacy of Docking Scores to Reflect Specificity

Docking software, which is widely used in computational drug design, typically relies on force field models to calculate the interaction energy between molecules and receptors(Ferreira et al., 2015; Lyu et al., 2023; Mysinger & Shoichet, 2010). These models focus on the joint binding energy of pocket-ligand pairs, denoted as $E(x, y)$. However, this approach does not fully address the specificity of binding interactions.

A significant drawback of force field models lies in their empirical design, which relies on training with existing molecular structures and properties (Zheng et al., 2022; Kitchen et al., 2004; Quiroga & Villarreal, 2016). This approach introduces biases, limiting the models' ability to accurately capture the full spectrum of molecular features and interactions. Consequently, when generative models produce molecules with structures that docking software erroneously favors, it leads to artificially inflated docking scores. This reveals a significant limitation of docking scores: they do not always accurately reflect the specific binding behavior of protein-molecule pairs. Instead, these scores can often reflect a bias towards certain molecular structures, misleadingly suggesting an unconditional binding ability (Eldridge et al., 1997; Wang et al., 2002).

3.2.2. Limitations in Modeling Specificity in Previous SBDD Methods

In the context of structure-based drug design (SBDD), the problem can be conceptualized as optimizing the generative model, denoted as θ, to maximize $p_\theta(x|y)$. Here, x represents the molecule, while y denotes the protein pocket. Advanced generative models are adept at producing molecules with high docking scores. However, these molecules often demonstrate a lack of selectivity, exhibiting high docking scores across a wide range of pockets within the test set.

This phenomenon can be further elucidated through the Bayesian formula:

$$p_\theta(x, y) = p_\theta(y|x)p_\theta(x). \tag{1}$$

We contend that the perceived progress made by contemporary advanced generative methods, evident in the enhanced $p_\theta(x, y)$, is primarily attributable to improvements in the unconditional components, $p_\theta(x)$. This development stands in stark contrast to the conditional component $p_\theta(y|x)$, which is crucial for achieving specificity in molecular binding. A significant repercussion of this disparity is the tendency of generated molecules to attain uniformly high docking scores throughout the test set, irrespective of the distinctiveness of the pocket binding sites. The ensuing experimental results, to be detailed in the section 5.1, reinforce our hypothesis.

4. Method

4.1. From Unconditional to Conditional

Our previous analysis underscores the need for a heightened focus on specific binding behaviors in SBDD to align with real-world requirements. In this section, we adopt a probabilistic framework to bridge the gap between the outputs of generative models and SBDD evaluation criteria. We use x to denote a molecule and y to denote a protein pocket from pre-defined finite pocket set $\mathbb{Y}$ and ligand set $\mathbb{X}$. Utilizing the Boltzmann distribution, the binding energy induces a joint probability describing the probability of complex formation:

$$p(x, y) = \frac{1}{Z} e^{-E(x,y)}, \tag{2}$$

where Z is the normalization coefficient.

Crucially, we focus on the conditional probability $p(y|x)$, the likelihood that molecule x binds specifically to pocket y among all potential targets. We propose to formulate a novel specificity metric by using the logarithm of the conditional probability, the larger the better:

$$\begin{aligned} \log p(y|x) &= \log \frac{p(x, y)}{p(x)} = \log \frac{\frac{1}{Z} e^{-E(x,y)}}{\sum_{y \in \mathbb{Y}} \frac{1}{Z} e^{-E(x,y)}} \\ &= -E(x, y) - \log \sum_{y \in \mathbb{Y}} e^{-E(x,y)}. \end{aligned} \tag{3}$$

Equation (3) indicates that to assess the specific binding ability, it is necessary to consider the docking scores of both positive pairs and negative pairs rather than solely focusing on the score of one pair of molecule and receptor. This approach allows us to quantitatively assess the specificity of molecule-pocket interactions from the pairwise interaction energy, highlighting a crucial factor for the efficacy of SBDD strategies.

4.2. Measure the specific binding

Inspired by the general specificity metric defined in the previous section, we propose the **Delta Score** to measure the generated model and reflect the specificity and efficacy of the generated molecules.

For a dataset with n pockets $y_1, y_2, \cdots, y_n$, the model to be evaluated generates m_i molecules $x_{i1}, x_{i2}, \cdots, x_{im_i}$ for each pocket y_i. Then for any $i \in \{1, \cdots, n\}$, $j \in$

$\{1, \cdots, m_i\}$, the specificity metric can be calculated by

$$S(x_{ij}, y_i) + \log \sum_{k=1}^{n} e^{-S(x_{ij}, y_k)} \tag{4}$$

$$\approx S(x_{ij}, y_i) - \min_{y_k} S(x_{ij}, y_k), \tag{5}$$

where we use a docking score, denoted as $S(x, y)$, to estimate the interaction energy $E(x, y)$. (4) is the negative of the metric defined in (3). (5) is a tight approximation of (4), with the log-sum-exp approximated by the max function. As a result, (5) has the same unit as the utilized docking score $S(x, y)$ and is also the smaller the better, analogous to existing docking scores.

Notice that in order to calculate the minimization term in (5) for all $i = 1 \cdots n$, $j = 1, \cdots m_i$, we need to perform docking for molecules with all possible pockets which is of quadratic complexity $O(n^2)$. To address this challenge, we provide a computational efficiency estimation of (5) by random sampling. Finally, we derive the Delta Score as the specificity metric of the generative model for a given pocket y_i:

$$\textbf{Delta Score}(y_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} (-S(x_{ij}, y_i) + S(x_{ij}, y_k)),$$
$$\text{for each } i, \text{we sample } k \in \{1, 2, \ldots, n\} \text{ with } k \neq i. \tag{6}$$

Intuitively, a smaller value of Delta Score indicates the molecules generated by the model specifically for target y_i, have a high affinity towards target y_i itself relative to the average affinity across all the other targets y_k. We provide more discussion of the derivation of Delta Score in Appendix A.

4.3. Generation with Specific binding

4.3.1. Overview of our method

We denote the molecule as $X = [c^X, v^X]$, where $[\cdot, \cdot]$ is the concatenation operator and $c \in R^{N_X \times 3}$ and $v \in R^{N_X \times K}$ denote atom Cartesian coordinates and one-hot atom types respectively, N_X is the number of atoms, K is the number of type of an atom respectively. Pockets are denoted as $Y = [c^Y, v^Y]$ similarly.

The overview of our method is shown in figure 3. Our primary objective is to enhance the specificity of the interaction between synthesized small molecules and their intended targets. At the core of our methodology, we develop a robust model that is adept at assessing this particular attribute. The trained model itself functions as an energy function, which plays a crucial role in guiding the diffusion-based synthesis of these molecules. By effectively utilizing this function, we can direct the generation of molecules towards those that demonstrate a heightened ability to bind specifically to their targets. A key requirement for the molecular generative process is that the probability $p_\theta(x_0|y)$ should be invariant to translation and rotation of the protein-ligand complex, which will be satisfied if the following properties hold (Guan et al., 2023a; Bao et al., 2022):

1. The initial distribution of diffusion process $p_\theta(X_T|Y)$ is an SE(3)-invariant distribution.

2. The transformation at each time step of the diffusion process is SE(3)-equivariant.

3. The energy function used to guide the diffusion process is invariant under SE(3) transformations.

To meet the above requirements, we move the center of mass (CoM) of the protein atoms to zero during initialization, and apply isotropic Gaussian noise as the coordinates of atoms at time step T, ensuring that the initial distribution is SE(3)-invariant. Furthermore, we adopt 3D-Equivariant graph neural network to parameterize the denoiser for molecular diffusion process as well as the encoder for the energy function. We provide detailed information on the feature updates of the SE(3)-Equivariant GNN network in Appendix B.

4.3.2. Molecular Diffusion Model

Following (Guan et al., 2023a), in the forward diffusion process, we define the transition kernels of the diffusion model for atomic coordinates and types using Gaussian noise and categorical noise, respectively. β_t is defined by fixed variance schedules, we denote $\alpha_t = 1 - \beta_t$, $\overline{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ and $\tilde{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$, we have:

$$q(X_t|X_{t-1}) = \prod_{i=1}^{N_X} \mathcal{N}(c^X_{t,i}; \sqrt{\alpha_t} c^X_{t-1,i}, \beta_t \mathcal{I}) \cdot \mathcal{C}(v^X_{t,i}|\alpha_t v^X_{t-1,i} + \beta_t/K) \tag{7}$$

$$q(X_t|X_0) = \prod_{i=1}^{N_X} \mathcal{N}(c^X_{t,i}; \sqrt{\overline{\alpha}_t} c^X_{0,i}, (1-\overline{\alpha}_t)\mathcal{I}) \cdot \mathcal{C}(v^X_{t,i}|\overline{\alpha}_t v^X_{0,i} + (1-\overline{\alpha}_t)/K) \tag{8}$$

The corresponding normal posterior if atom coordinates and categorical posterior of atom types can be computed as:

$$q(X_{t-1}|X_t, X_0) = \prod_{i=1}^{N_X} \mathcal{N}(c^X_{t-1,i}; \tilde{\mu}(c^X_{t,i}, c^X_{0,i}), \tilde{\beta}_t \mathcal{I}) \cdot \mathcal{C}(v^X_{t-1,i}|\tilde{c}(v^X_{t,i}, v^X_{0,i})). \tag{9}$$

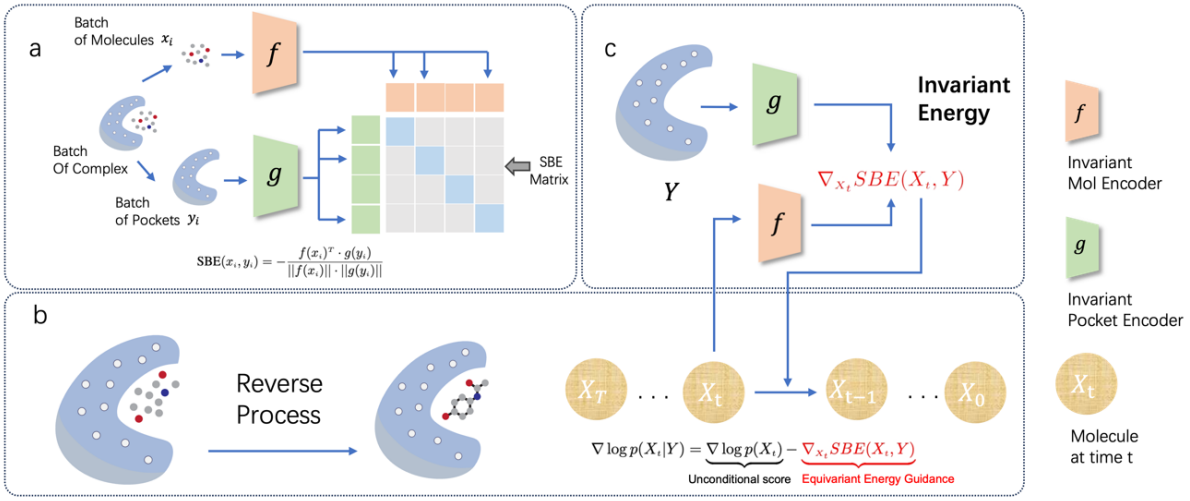


Figure 3. The main framework of our SBE-Diff model. Part a) depicts the training phase of the Specific Binding Energy (SBE) model, where in-batch softmax loss is employed. We use ligands designated for different pockets as negative instances for the current pocket, aiming to address the specific binding. Part b) describes the model's reverse diffusion process, wherein the transition from X_{t+1} to X_t is guided by minimizing the specific binding energy. Part c) demonstrates the calculation of the SBE. This is achieved by utilizing pocket and molecule embeddings, produced by their respective specially trained encoders.

$$\begin{aligned}\tilde{\mu}(c^X_{t,i}, c^X_{0,i}) &= \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t} c^X_{0,i} + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t} c^X_{t,i}, \\ c^*(v^X_{t,i}, v^X_{0,i}) &= [\overline{\alpha}_{t-1} v^X_{0,i} + \frac{(1-\overline{\alpha}_{t-1})}{K}] \\ &\quad \odot [\alpha_t v^X_{t,i} + \frac{(1-\alpha_t)}{K}], \\ \tilde{c}(v^X_{t,i}, v^X_{0,i}) &= \frac{c^*}{\sum_{k=1}^K c^*_k}. \end{aligned} \tag{10}$$

During the generative process, we will recover molecule X_0 from the initial noise X_T. We predict the atom coordinates $\hat{c}^X_{0|t,i}$ and atom types $\hat{v}^X_{0|t,i}$ using denoising SE(3)-Equivariant network θ to approximate the reverse distribution at every time step t as follows:

$$\begin{aligned} p_\theta(X_{t-1}|X_t) = \prod_{i=1}^{N_X} &\mathcal{N}(c^X_{t-1,i}; \tilde{\mu}(c^X_{t,i}, \hat{c}^X_{0|t,i}), \tilde{\beta}_t \mathcal{I}) \\ &\cdot \mathcal{C}(v^X_{t-1,i} | \tilde{c}(v^X_{t,i}, \hat{v}^X_{0|t,i})). \end{aligned} \tag{11}$$

We also provide detailed information on the parameterization of the diffusion process in Appendix B.

4.3.3. Specific Binding Energy Function

Then we introduce the SBE-model we use to serve as the guided energy function. The SBE-model consists of a molecule encoder SBE-Enc$_X$ as well as a pocket encoder SBE-Enc$_Y$. For a protein pocket y_i and a molecule x_i, we first define the specific binding energy (SBE) as:

$$\begin{aligned} &\mathrm{SBE}(x_i, y_i) = -\frac{f(x_i)^T \cdot g(y_i)}{||f(x_i)|| \cdot ||g(y_i)||}, \text{ where} \\ &f(x_i) = \text{SBE-Enc}_X(X_i), g(y_i) = \text{SBE-Enc}_Y(Y_i). \end{aligned} \tag{12}$$

The smaller the SBE is, the stronger the binding specificity. In order to provide guidance for the generative process throughout all diffusion steps, we train this energy function to be time-dependent, which receives the noisy molecule data $x_{t,i}$ as inputs. During the training phase, we randomly sample time step t and apply formula 8 to sample the noisy molecule data $x_{t,i}$ from ground truth x_i. For a batch of size N, we have pairwise data $\{(x_i, y_i)\}_{i=1}^N$. we extract a list of protein pockets $\{y_i\}_{i=1}^N$ and a list of corresponding molecules $\{x_i\}_{i=1}^N$. Combining them together results in N^2 pairs (x_i, y_j) where $i, j \in [1, N]$. When $i = j$ it is a positive pair, and when $i \neq j$ it is a negative pair.

Our first training goal is to identify the true binding molecules before other molecules for a given protein pocket:

$$\mathcal{L}_p(y_i, \{x_{t,j}\}_{j=1}^N) = -\frac{1}{N} \log \frac{\exp(-\mathrm{SBE}(x_{t,i}, y_i)/\tau)}{\sum_j \exp(-\mathrm{SBE}(x_{t,j}, y_i)/\tau)}, \tag{13}$$

Symmetrically, our second training goal is to identify the true binders from a batch of protein pockets.

$$\mathcal{L}_m(x_{i,t}, \{y_j\}_{j=1}^N) = -\frac{1}{N} \log \frac{\exp(-\mathrm{SBE}(x_{t,i}, y_i)/\tau)}{\sum_j \exp(-\mathrm{SBE}(x_{t,i}, y_j)/\tau)}, \tag{14}$$

From the theoretical side, our training loss is exactly the same format as the log conditional probability in Section 4.1. Minimizing the Pocket to Mol loss $\mathcal{L}_p$ is equivalent to maximizing $p(y|x)$, the probability that a molecule binds to its corresponding protein pocket specifically. Similarly, minimizing the Mol to Pocket loss $\mathcal{L}_m$ is equivalent to maximizing $p(x|y)$, the probability that a pocket binds to its corresponding ligand specifically.

4.3.4. SB-ENERGY MINIMIZER DIFFUSION MODEL

We propose an energy-guided diffusion model that dynamically adjusts the types and positions of atoms in the generative process, leveraging the SBE-model. We minimize this SB-energy using the techniques of the Energy-Guided Diffusion Model (Bao et al., 2022) to facilitate the ability of the generated molecule to specifically bind to its target:

$$p(X|Y) \propto p(X) \cdot \exp\left(-\mathcal{E}(X,Y)\right), \tag{15}$$

where we adopt SBE as the guided energy function $\mathcal{E}(X,Y)$. More concretely, at each time step of the reverse diffusion process, we feed the noisy molecule X_t and its corresponding pocket Y into the SBE-model to calculate the SB-Energy and update the denoising process using its gradient with respect to the molecule $\nabla_{X_t}\text{SBE}(X_t,Y)$:

$$\begin{aligned}\hat{p}_\theta(X_{t-1}|X_t,Y) = \\ \prod_{i=1}^{N_X}\mathcal{N}(c^X_{t-1,i};\tilde{\mu}(c^X_{t,i},\hat{c}^X_{0|t,i})-w_1\nabla_{c^X_t}\text{SBE}(X_t,Y),\tilde{\beta}_t\mathcal{I}) \\ \cdot\mathcal{C}(v^X_{t-1,i}|\tilde{c}(v^X_{t,i},\hat{v}^X_{0|t,i})-w_2\nabla_{v^X_t}\text{SBE}(X_t,Y)).\end{aligned} \tag{16}$$

where w_1 and w_2 are empirical coefficients used to adjust the intensity of the energy condition, we set w_1 to 0.1 and w_2 to 1e-4 in practical. The complete sampling procedure is shown in Algorithm 1.

Algorithm 1 Sampling Procedure of SBE-Diff

Input: The pocket Y, the model ϕ_θ, coefficients w_1 w_2.
Output: Generated ligand molecule X that binds to the protein pocket specifically.

1: Sample the number of atoms N_X based on a prior distribution conditioned on the pocket size.
2: Model the whole complex as a graph $[X,Y]$.
3: Move center of mass of protein atoms to zero.
4: Sample initial molecular atom coordinates c_T and atom types v_T: $c_T \in \mathcal{N}(0,I)$, v_T = one hot$(\arg\max_i g_i)$, where $g \sim \text{Gumbel}(0,1)$.
5: **for** t in $T, T-1, \ldots, 1$ **do**
6: Predict $\hat{X}_0 = [\hat{c}^X_0, \hat{v}^X_0]$ from c^X_t, v^X_t with ϕ_θ: $\hat{c}^X_0, \hat{v}^X_0 = \phi_\theta(c^X_t, v^X_t, t, Y)$.
7: Calculate the posterior $\hat{p}_\theta(X_{t-1}|X_t,Y)$ according to equation 16.
8: Sample c^X_{t-1} and v^X_{t-1} from the posterior $\hat{p}_\theta$.
9: **end for**

5. Experiments

5.1. Main Results

In our study, we conducted experiments on the CrossDocked2020 dataset, following the data split outlined by Guan et al. (2023a). We use the CrossDocked2020 dataset to train both the generative model and the SBE model.

Given the necessity of performing docking after shuffling binding sites, we faced a unique challenge: using the original generated 3D conformations of molecules would be unfair and unreasonable, as a molecule created for one pocket should be docked to a different pocket for this experiment. To address this, we converted all generated molecules into 1d/2d representations(Rdkit Mols) and let Glide to generate initial conformations. These conformations were then docked, and the one with the highest docking score was selected. Beyond the conventional docking score, we introduce the delta score, as detailed in formula 6. In order to calculate the delta score, molecules generated for each target are docked with another randomly sampled target. For a fair comparison, we have fixed the random seeds. We show more results in Appendix D. Additionally, we present the ratio indicating instances where the generated molecule outperforms the reference ligand for a specific pocket. In this context, 'better performance' is defined as the scenario where the generated molecule achieves both a higher absolute docking score and a better delta score compared to the corresponding reference ligand. For the aggregation of different binding sites, we show the results in mean and median values.

Table 2. Experimental Results for different methods. We show the absolute docking score, delta score, and ratio of the generated molecule better than the reference ligand. For each metric, the best result is **bold** and the second best result is underlined.

	Absolute ↑	Delta ↑		Ratio ↑	
Methods	mean	mean	median	mean	median
trainset	5.727	0.044	-0.073	0.159	0.053
AR	5.737	0.393	0.200	0.150	0.039
Pocket2Mol	5.946	0.437	0.106	0.170	0.050
DrugGPS	6.554	<u>0.490</u>	**0.387**	0.235	0.134
TargetDiff	6.665	0.335	0.102	0.259	0.129
DecompDiff	**7.102**	0.354	0.220	<u>0.274</u>	<u>0.133</u>
SBE-Diff	<u>6.815</u>	**0.552**	<u>0.250</u>	**0.300**	**0.216**
Reference	6.632	1.158	1.029	-	-

Table 2 and Figure 1 collectively highlight a critical trend in structure-based drug design (SBDD). While several methods, including recent ones like Targetdiff and DecompDiff, achieve absolute docking scores surpassing those of reference ligands, they fall short in delta scores, indicating a deficiency in specific binding ability. This discrepancy is particularly pronounced when compared with older methods like AR and Pocket2Mol. The chronological plot in Figure 1 further illustrates this point; despite a clear upward trajectory in docking scores over time, delta scores, represented by the green line, fluctuate without significant improvement and remain substantially lower than those of reference ligands. This observation underscores our argument that **current advancements in SBDD are primarily focused on improving unconditional part ($p_\theta(x)$), rather than the more crucial conditional aspects of molecular binding ($p_\theta(y|x)$).**

Table 3. Comparison of SBE-Diff with TargetDiff in terms of different metrics commonly used in SBDD evaluation. Difference is shown in percentage.

	ori score ↑	shuffle score ↓	delta score ↑	Ratio ↑	QED ↑	SA ↑	Diversity ↑
TargetDiff	6.665	6.320	0.335	0.259	0.503	0.587	0.880
SBE-Diff	6.815	6.261	0.552	0.300	0.497	0.586	0.879
Difference	2.2%	0.9%	64.8%	15.8%	-1.2%	-0.0%	-0.0%

As shown in table 2, our SBE-Diff approach stands out remarkably in the comparative analysis, consistently ranking either first or second across all evaluated metrics, as detailed in our results. This performance is particularly noteworthy in terms of Delta Scores and the ratio of surpassing reference ligands, where our method demonstrates exceptional proficiency. Such results unequivocally indicate that the SBE-Diff approach excels in directing the generative process towards enhanced specificity in molecular binding.

A notable observation from our study is the better performance of DecompDiff over TargetDiff in terms of conditional binding effectiveness. This outcome is likely attributed to DecompDiff's innovative approach of utilizing pocket information to establish priors for atom clustering, thereby explicitly and effectively modeling conditional binding. Similarly, DrugGPS demonstrates remarkable capabilities in specific binding scenarios. This is largely due to its unique strategy of constructing a subpocket prototype-molecular motif interaction graph, which directly and proficiently models the conditional binding aspect. These findings underscore the importance of tailored approaches in enhancing conditional binding effectiveness in drug design.

5.2. Ablation and Analysis

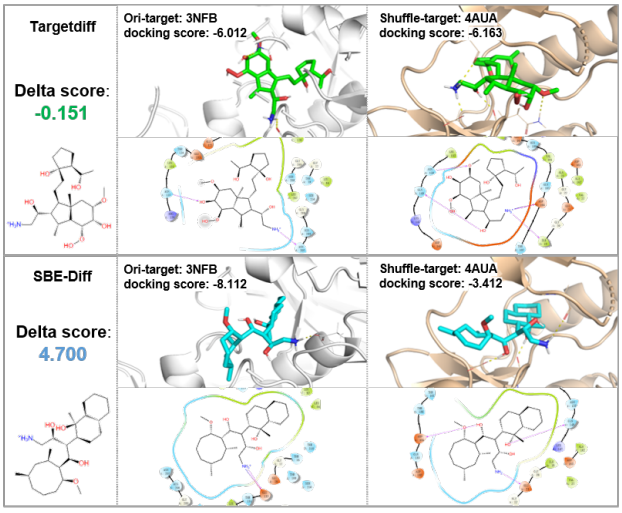


Figure 4. Case study of generated molecules docked with its original target and shuffled target.

5.2.1. Effectiveness of the SBE guidance

Our diffusion process and model bear similarities to Targetdiff, with a notable distinction being our Specific Binding Energy (SBE) guidance component. To demonstrate the effectiveness of our SBE guidance framework, we conducted a comparative analysis with Targetdiff. Alongside docking scores, we included other common SBDD benchmarking metrics in our evaluation.

The results, detailed in Table 3, show that our SBE-Diff approach positively impacts specificity without compromising other key metrics like QED, Synthesizability(SA), and diversity. Notably, our method enhances the absolute docking score against the specific target and reduces the score post-shuffling. This leads to a significant increase in the delta score, effectively capturing the improvement in specificity.

5.2.2. Visualization

A case of the molecule generated by SBE-Diff versus the molecule generated by TargetDiff is shown in figure 4. The SBE-Diff molecule achieves a significantly higher Delta score of 4.700, in contrast to the Targetdiff molecule with a modest Delta score of -0.151. The Targetdiff molecule exhibits similar docking scores on both ori-target and shuffle-target. This observation aligns with our understanding that molecules rich in hydroxyl groups, e.g. sugars and polyols, can bind to various protein pockets through numerous hydrogen bonds and other interaction mechanisms. SBE-Diff performs well, achieving a docking score of -8.112 on ori-target. SBE-Diff does this by reducing these unnecessary hydroxyl groups while adding flexible hydrophobic moiety to the small molecule, increasing the contribution of the entropy effect of the small molecule binding process. Interestingly, SBE-Diff also removes the problematic peroxide group present in the original molecule. SBE-Diff helps increase molecule specificity by reducing potential affinity for the shuffle-target, while also improving the rationality of binding pose to the intended target. We show more cases in appendix C.

6. Conclusion

In our study, we analyze and rethink the current state of structure-based drug design, identifying a significant focus on generating molecules with high docking scores that primarily address unconditional aspects. We note a considerable lack of emphasis on specificity, which is crucial in

effective drug design. To tackle this issue, we introduce a new evaluation metric and a guidance framework centered on conditional binding to enhance specificity. Our experimental results validate our observation about the field's inclination towards unconditional aspects and demonstrate the effectiveness of our proposed methods in improving specificity. Our work not only advocates for a paradigm shift in the approach to drug design but also provides a concrete and effective direction for achieving this goal.

Impact Statement This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgements This work is supported by the National Key R&D Program of China No.2021YFF1201600, Beijing Frontier Research Center for Biological Structure, and Beijing Academy of Artificial Intelligence (BAAI).

References

Bao, F., Zhao, M., Hao, Z., Li, P., Li, C., and Zhu, J. Equivariant energy-guided sde for inverse molecular design. *arXiv preprint arXiv:2209.15408*, 2022.

Buttenschoen, M., Morris, G. M., and Deane, C. M. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 2024.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V., and Mee, R. P. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11:425–445, 1997.

Ferreira, L. G., Dos Santos, R. N., Oliva, G., and Andricopulo, A. D. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. 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.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

Gao, B., Qiang, B., Tan, H., Jia, Y., Ren, M., Lu, M., Liu, J., Ma, W.-Y., and Lan, Y. DrugCLIP: Contrasive protein-molecule representation learning for virtual screening. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=lAbCgNcxm7.

Geiger, M. and Smidt, T. e3nn: Euclidean neural networks, 2022.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations*, 2023a.

Guan, J., Zhou, X., Yang, Y., Bao, Y., Peng, J., Ma, J., Liu, Q., Wang, L., and Gu, Q. Decompdiff: Diffusion models with decomposed priors for structure-based drug design. 2023b.

Harris, C., Didi, K., Jamasb, A. R., Joshi, C. K., Mathis, S. V., Lio, P., and Blundell, T. Benchmarking generated poses: How rational is structure-based drug design with generative models? *arXiv preprint arXiv:2308.07413*, 2023.

Harrison, R. K. Phase ii and phase iii failures: 2013–2015. *Nat Rev Drug Discov*, 15(12):817–818, 2016.

Hoogeboom, E., Satorras, V. G., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

Lin, A., Giuliano, C., Palladino, A., John, K., Abramowicz, C., Yuan, M., Sausville, E., Lukow, D., Liu, L., Chait, A., et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. sci. transl. med. 11: eaaw8412, 2019.

Long, S., Zhou, Y., Dai, X., and Zhou, H. Zero-shot 3d drug design by sketching and generating. In *NeurIPS*, 2022.

Luo, S., Guan, J., Ma, J., and Peng, J. A 3d generative model for structure-based drug design. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6229–6239. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf.

Lyu, J., Irwin, J. J., and Shoichet, B. K. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*, pp. 1–7, 2023.

Masuda, T., Ragoza, M., and Koes, D. R. Generating 3d molecular structures conditional on a receptor binding site with deep generative models, 2020.

Mysinger, M. M. and Shoichet, B. K. Rapid context-dependent ligand desolvation in molecular docking. *Journal of chemical information and modeling*, 50(9):1561–1573, 2010.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, 2022.

Quiroga, R. and Villarreal, M. A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Satorras, V. G., Hoogeboom, E., and Welling, M. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Schneider, P. and Schneider, G. De novo design at the edge of chaos: Miniperspective. *Journal of medicinal chemistry*, 59(9):4077–4086, 2016.

Trott, O. and Olson, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

Wang, R., Lai, L., and Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design*, 16:11–26, 2002.

Wong, C. H., Siah, K. W., and Lo, A. W. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

Zhang, X., Gao, H., Wang, H., Chen, Z., Zhang, Z., Chen, X., Li, Y., Qi, Y., and Wang, R. Planet: A multi-objective graph neural network model for protein–ligand binding affinity prediction. *Journal of Chemical Information and Modeling*, 2023.

Zhang, Z. and Liu, Q. Learning subpocket prototypes for generalizable structure-based drug design. *ICML*, 2023.

Zhao, M., Bao, F., Li, C., and Zhu, J. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.

Zheng, L., Meng, J., Jiang, K., Lan, H., Wang, Z., Lin, M., Li, W., Guo, H., Wei, Y., and Mu, Y. Improving protein–ligand docking and screening accuracies by incorporating a scoring function correction term. *Briefings in Bioinformatics*, 23(3):bbac051, 2022.

A. Delta Score

A.1. General Specificity Metric

Existing docking scores typically calculate the interaction energy between the molecule and the protein pocket to reflect the binding strength (Friesner et al., 2004). We denote the interaction energy as $E(x, y)$, with the molecule and the pocket denoted as x and y, respectively. Utilizing the Boltzmann distribution, the energy naturally defines a joint probability describing the probability that x and y can bind with each other.

$$p(x, y) = \frac{1}{Z} e^{-E(x,y)}, \tag{17}$$

where $Z = \sum_{x \in \mathbb{X}} \sum_{y \in \mathbb{Y}} e^{-E(x,y)}$ is the normalization coefficient and the sum is over the pre-defined finite pocket set $\mathbb{Y}$ and ligand set $\mathbb{X}$. As a result, existing docking scores solely describe the joint probability of pocket-ligand pairs and are unable to reflect the specificity of binding interactions $p(y|x)$, which describes the probability that molecule x binds specifically to pocket y among all potential targets.

To address the challenge that existing docking scores overlook binding specificity, we propose to focus on the conditional probability $p(y|x)$ and formulate a novel specificity metric. Consider the logarithm of the specific binding probability for a molecule x and a pocket y:

$$\log p(y|x) = \log \frac{p(x, y)}{p(x)} = \log \frac{\frac{1}{Z} e^{-E(x,y)}}{\sum_{y \in \mathbb{Y}} \frac{1}{Z} e^{-E(x,y)}} = -E(x, y) - \log \sum_{y \in \mathbb{Y}} e^{-E(x,y)}. \tag{18}$$

Therefore, (18) functions as a specificity metric of binding and can be calculated from the pairwise interaction energy.

A.2. Delta Score: A Specificity Metric for Generated Models

In SBDD benchmarking, conventional methods use docking scores to measure the quality of the generated models and which fails to explicitly embody the specific binding behavior. Inspired by the general specificity metric defined in the previous section, we propose the Delta Score to measure the generated model and reflect the specificity and efficacy of the generated molecules.

For a dataset with n pockets $y_1, y_2, \cdots, y_n$, the model generates m_i molecules $x_{i1}, x_{i2}, \cdots, x_{im_i}$ for each pocket y_i. Then for any $i \in \{1, \cdots, n\}$, $j \in \{1, \cdots, m_i\}$, specificity metric can be calculated by

$$-S(x_{ij}, y_i) - \log \sum_{k=1}^{n} e^{-S(x_{ij}, y_k)}, \tag{19}$$

where we use a docking score, denoted as $S(x, y)$, to estimate the interaction energy $E(x, y)$.

Next, we provide a simpler form of the specificity metric by using the following lemma.

Lemma A.1 (Log-sum-exp approximation)**.** *For a finite set of real numbers $z_1, z_2, \cdots, z_N$, the log-sum-exp function can be upper and lower bounded by the maximum function.*

$$\max\{z_1, \cdots, z_N\} \leq \log(\sum_{k=1}^{N} \exp(z_k)) \leq \max\{z_1, \cdots, z_N\} + \log(N). \tag{20}$$

Proof.

$$\begin{aligned} \max\{z_1, \cdots, z_N\} &= \log(\exp(\max\{z_1, \cdots, z_N\})) \\ &\leq \log(\exp(z_1) +, \cdots, \exp(z_1))) \\ &\leq \log(N \exp(\max\{z_1, \cdots, z_N\})) \\ &= \max\{z_1, \cdots, z_N\} + \log(N). \end{aligned} \tag{21}$$

□

Therefore, we can utilize the maximum function to approximate the log-sum-exp function. The approximation error is small when the exponential of the largest element is much larger than the exponential of other elements. It is the case for

docking scores since they are negative in most cases. Then (19) can be approximated by $-S(x_{ij}, y_i) - \max_{y_k}(-S(x_{ij}, y_k))$. Analogous to existing docking scores, we define the initial Delta Score as

$$S(x_{ij}, y_i) - \min_{y_k} S(x_{ij}, y_k), \tag{22}$$

which has the same unit as the utilized docking score $S(x, y)$ and is also the smaller the better. Notice that to calculate (22) for all $i = 1 \cdots n$, $j = 1, \cdots m_i$, we need to perform docking for molecules with all possible pockets which is of quadratic complexity $O(n^2)$. To address this challenge, we provide a computationally efficient estimation of the initial Delta Score

$$S(x_{ij}, y_i) - S(x_{ij}, y_k), \text{where } k \text{ is randomly sampled in } \{1, 2, \ldots, n\} \text{ with } k \neq i. \tag{23}$$

Finally, we arrive at the definition of the Delta Score for a given pocket: For any pocket y_i, we sample $k \in \{1, 2, \ldots, n\}$ with $k \neq i$:

$$\textbf{Delta Score}(y_i) \triangleq \frac{1}{m_i} \sum_{j=1}^{m_i} (S(x_{ij}, y_i) - S(x_{ij}, y_k)). \tag{24}$$

In fact, (23) serves as an unbiased estimation of the quantity $S(x_{ij}, y_i) - \frac{1}{n-1}\sum_{k \neq i} S(x_{ij}, y_k) \leq S(x_{ij}, y_i) - \min_{y_k} S(x_{ij}, y_k)$, consistently yielding lower values than (22). Both formulations can be intuitively explained by emphasizing the binding affinity of molecules with their specific target protein pockets, as opposed to multiple similar pockets without selectivity. A smaller value of (22) indicates the molecule x_{ij}, generated by the model specifically for target y_i, has a high affinity towards target $y_i$itself, relative to any alternative target y_k. Whereas (23) relaxes this constraint, allowing for comparisons of affinity relative to the average affinity across all other targets y_k.

B. SE(3)-Equivariant GNN

In Section 4, we denote molecules and pockets as $X = [c^X, v^X]$ and $Y = [c^Y, v^Y]$ respectively.

B.1. Parameterization of diffuison process

Denote the GNN network as ϕ_θ. We model the whole complex as a graph $[X, Y]$ and let the neural network predict $[x_0, v_0]$ at each time step:

$$[\hat{x_0}, \hat{v_0}] = \phi_\theta(X_t, t, Y) = \phi_\theta([c_t^X, v_t^X], t, Y). \tag{25}$$

At the l-th layer, the atom hidden embedding $\mathbf{h}$ and coordinates c are updated alternately as follows:

$$\begin{aligned} \mathbf{h}i^{l+1} &= \mathbf{h}i^l + \sum_{j \in \mathcal{V}, j \neq i} f_h(d_{ij}^l, \mathbf{h}i^l, \mathbf{h}j^l, eij; \theta_h), \\ c_i^{l+1} &= c_i^l + \sum_{j \in \mathcal{V}, j \neq i} (c_i^l - c_j^l) f_c(d_{ij}^l, \mathbf{h}i^{l+1}, \mathbf{h}j^{l+1}, eij; \theta_x) \cdot \nVdash_{mol}. \end{aligned} \tag{26}$$

, where d_{ij} is the euclidean distance between atoms i and j along with an additional feature e_{ij} indicating the connection type (protein-protein, ligand-ligand, or protein-ligand). The mask $\nVdash_{mol}$ is applied to prevent updates on protein atom coordinates. The initial atom hidden embedding $\mathbf{h}_0$ is obtained through a linear embedding layer that encodes atom information. The final atom hidden embedding $\mathbf{h}_L$ is then passed through a MLP and a softmax function to obtain the predicted value $\hat{v_0}$.

B.2. Parameterization of SBE-Encoder

We also adopt 3D-Equivariant graph neural network to parameterize SBE-Enc$_X$ and SBE-Enc$_Y$. Different from B.1, molecules and pockets are encoded separately instead of forming a single full graph. Taking SBE-Enc$_X$ as an example, SBE-Enc$_Y$ follows the same principle.

At the l-th layer, we do not update the atom coordinates but only the hidden embedding $\mathbf{h}$:

$$\mathbf{h}i^{l+1} = \mathbf{h}i^l + \sum_{j \in \mathcal{V}, j \neq i} f_h(d_{ij}^l, \mathbf{h}i^l, \mathbf{h}j^l, eij; \theta_h). \tag{27}$$

In the final layer, we obtain the encoding of the entire molecule by performing average pooling over all the atoms:

$$f(x) = \text{SBE-Enc}_x(X) = \frac{1}{N_X}\sum_{i=1}^{N_X}\mathbf{h}_i^L. \tag{28}$$

C. Visualizations

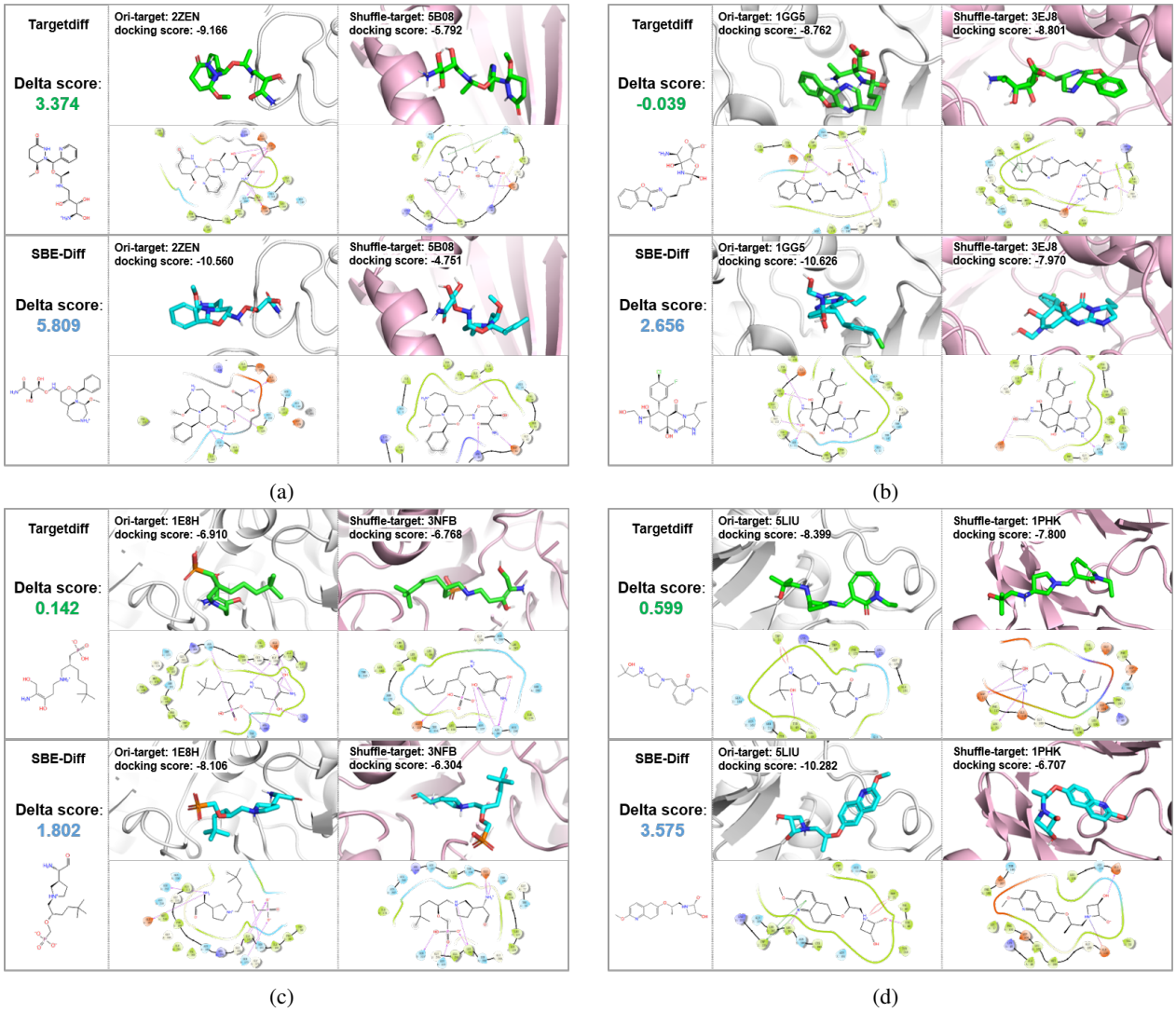


Figure 5. Visualizations of molecules generated by SBE-Diff versus molecules generated by TargetDiff. Molecules are docked with their original targets and shuffled targets.

D. More Results of Delta Score

To compare between different methods, we use the same seed for sampling/shuffling the pockets. To make the results more reliable, we add experiments based on 5 different seeds for shuffling pockets, and report average delta scores and corresponding standard deviation, as shown in Table 4. It is worth noting that, when averaging all pockets in the test set, the delta score is actually an unbiased estimate of the difference between the docking scores of molecules with their targets and the average docking scores with other pockets, as shown in Appendix A.2. Therefore, from a theoretical perspective, it is relatively stable. Practically, the results shown above prove the stability of our proposed metric.

Table 4. Delta scores of different random seed.

Method	**Seed 2023**	**Seed 2022**	**Seed 2021**	**Seed 2020**	**Seed 2019**	**Mean/Std**
Ground Truth	1.158	1.126	1.150	1.062	1.254	1.150 ± 0.062
SBE-Diff	0.552	0.652	0.629	0.464	0.608	0.581 ± 0.067
TargetDiff	0.335	0.405	0.365	0.315	0.400	0.364 ± 0.035
Pocket2Mol	0.437	0.381	0.347	0.451	0.384	0.400 ± 0.038
DecompDiff	0.354	0.425	0.372	0.328	0.429	0.382 ± 0.040