Delta Score: Improving the Binding Assessment of Structure-Based Drug Design Methods

Minsi Ren[1], Bowen Gao[2] , Bo Qiang[3] , Yanyan Lan[2]*
[1]Institute of Automation, Chinese Academy of Sciences
[2]Institute for AI Industry Research (AIR), Tsinghua University
[3]Department of Pharmaceutical Science, Peking University

Abstract

Structure-based drug design (SBDD) stands at the forefront of drug discovery, focusing on developing molecules that target specific binding pockets. Recent advances in this area have witnessed the adoption of deep generative models, modeling SBDD as a conditional generation task where the target structure serves as context. Despite previous claims that generated ligands outperform their respective ground truth counterparts in terms of docking score evaluation, our analysis reveals that these perceived performance improvements are attributed to inherent biases within the scoring systems themselves, rather than an accurate assessment of the ligands' binding affinity. To address this issue, we introduce the delta score, a new evaluation metric emphasizing docking scores that prioritize specificity. Our experiments reveal that molecules produced by current deep generative models significantly lag behind ground truth reference ligands when assessed with the delta score. This novel metric not only complements existing benchmarks but also provides a pivotal direction for subsequent research in the domain.

1 Introduction

In the field of drug discovery, the development of novel small molecules that could form stable binding complexes with a specific disease-related target, known as structure-based drug design (SBDD) tasks, is of paramount importance. With the availability of an increasing amount of structural data, such as the PDBbind[21] and Crossdocked dataset[2], many deep generative models [12, 17, 11] have been proposed to address SBDD by formulating it as a target-based conditional generation task, resulting in remarkable progress.

Specifically, the assessment of these models has predominantly focused on docking scores between generated molecules and designated targets, using docking software such as Vina [19]. Despite the claims made by state-of-the-art models that a majority of the ligands they generate outperform the docking score of ground truth ligands in test sets [4], it raises the questions: Has the issue of 3D molecule generation been conclusively resolved? Do these scores align with real-world biological needs?

To address these inquiries, we conducted an experiment on the CrossDocked2020 dataset to achieve a fair comparison between various deep generative models. Our reproduced outcomes have exhibited evident enhancements of these deep generative models compared to the ground-truth, at least for a certain proportion of targets, which are consistent with previous claims. However, surprisingly, we discovered that a basic random sampling approach from Zinc also produced superior outcomes when compared to the ground-truth. This exceptional result prompts us to reevaluate the chosen evaluation metric.

*Correspondence to lanyanyan@air.tsinghua.edu.cn

Preprint. Under review.

Our analysis show the shortcomings of using “docking score” as the primary evaluation metric: the overall averaged score could be artificially inflated by characterizing certain biases, without truly capturing the matching degree between the molecules and the target. Therefore, we propose a complementary metric named "delta score", which allows us to distinguish specificity in binding to the target by subtracting the average matching degree, thereby better accommodating the requirements of applications in structure-based drug molecule design.

2 Related Work

With the emergence of geometric models [18, 3], 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 [15], auto-regressive models [13, 17, 10], and diffusion models [4, 5]. Most of these works used Vina score [19] for binding affinity evaluation, which is a typical docking software to predict the interactions between a small molecule and a protein target, similar to Glide [6] and Gold [20]. However, there is limited discussion on the suitability of this widely used docking metric in assessment of SBDD methods.

3 Experimental Analysis w.r.t. Docking Score

We first conduct experiments on CrossDocked2020 dataset [2] to compare different deep generative models, including auto-regressive model[12] (denoted as AR), Shape2mol[11], Pocket2mol[17] and Targetdiff[4]. In addition, we randomly sample small molecules for each target from Zinc dataset [8], to form a baseline named Random_Zinc method. The data preprocessing and splitting are all following Luo et al. [12] to ensure a fair comparison. For each method, we generate 100 molecules for each target pocket in the test set and calculate the affinity score using Glide instead of Vina, since Glide has demonstrated superior performance in predicting both accurate conformation and binding affinity [22]. It is important to note that the Vina evaluation results display a high degree of similarity, therefore the conclusions drawn can be generalized across different docking scores.

3.1 Experimental Results

From the results in Table 1, we find that: 1) there is relatively little difference in the averaged docking scores generated by different models, including ground truth and Random_Zinc; 2) Targetdiff consistently outperforms ground truth; 3) each method is capable of generating molecules that outperform ground truth performance on different targets, for example, Targetdiff outperforms ground truth on nearly half targets and even Random_Zinc has the ability to outperform ground truth on 22.4% targets. Evidently, these findings have sparked concerns regarding the reliability of the docking metrics utilized.

Table 1: Glide Docking Scores

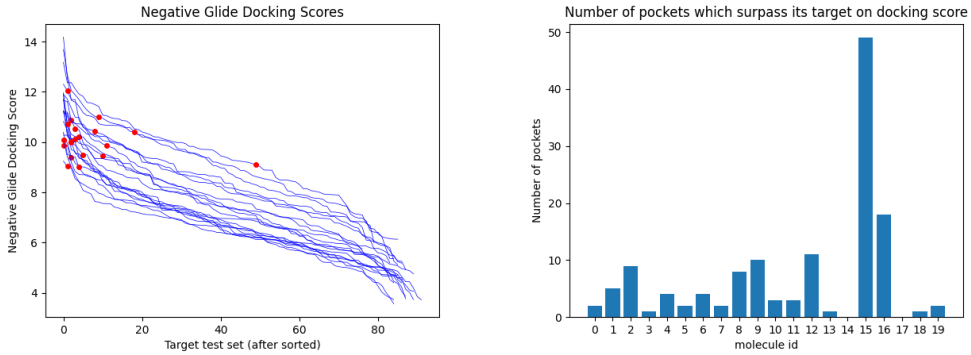
Dataset	CrossDocked 2020		
Methods	**Mean of mean (↓)**	**Median of mean (↓)**	**Better than GT (↑)**
Ground Truth	-6.367	-6.581	-
AR	-5.833	-5.666	0.359
Pocket2mol	-6.282	-6.170	0.382
Shape2mol	-5.631	-5.663	0.265
Targetdiff	-6.670	-6.742	0.489
Random_Zinc	-5.543	-5.564	0.224

3.2 Case Study

Upon further analysis of the small molecules generated by Targetdiff, particularly those exhibiting favorable docking scores, we have discovered some intriguing properties:

- These molecules typically exhibit highly intricate cyclic structures, implying their extremely low likelihood of existing in reality and the significant challenges to synthesize.
- These molecules typically have promising docking scores against most of pockets in the test set, rather than only their specific target, which means they may be pan-assay interference compounds (PAINS) or have poor specificity.

More specifically, we randomly select 20 pockets in the test set and generate 100 small molecules for each pocket. We select the one with the best docking score on each pocket, totaling 20 small molecules. All these small molecules have a docking score of over -9 on their true targets. We cross dock them with all the pockets in the test set, the results are shown as Figure 1. We notice that out of the 20 small molecules, only 2 have a top 1 docking score for their true targets. Furthermore, we have observed that for almost a quarter of the molecules, the docking scores are higher for over 10 other pockets as compared to their true targets. For a more detailed demonstration, we choose to display one generated molecule with a glide docking score of up to -10.111 against its target pocket in Appendix B.



(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 1: In the left image, each line represents the sorted docking score of a small molecule against all pockets in the test set. 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.

3.3 Analysis

Based on previous experimental results and case studies, we have concluded that current docking scores have limitations in accurately evaluating the binding relationships between generated molecules and targets, leading to false positive issues.

According to [14, 1, 16], docking software generally employs force field models to assess the interaction energy between molecules and receptors. However, these models are often empirical and trained on known structures and properties, resulting in inherent limitations. They may fail to accurately capture all molecular features and types of interactions, leading to biases. One consequence of this bias is that docking software may assign high scores to small molecules with some certain specific structures, such as PAINS, even if they are actually false positives.

While docking scores can offer insights into the binding affinity between small molecules and target proteins, they may not provide a complete assessment of selectivity. It is crucial to consider the off-target effects, which arise when a drug molecule interacts with unintended targets, leading to potential adverse reactions and impacting the overall therapeutic outcome [7, 23, 9].

To develop an metric that can effectively counteract bias, a straightforward yet effective way is to consider the difference in scores between positive pairs and negative pairs rather than solely focusing on the score of one pair of molecule and receptor. Based on this, we introduce delta score in Section 4.1. By incorporating the variation in affinity scores across different targets, this metric can also assess the specific binding capacity of a small molecule to its target.

4 Delta Score

4.1 Definition

Suppose the test set contains n target pockets $p_1, p_2...p_i..., p_n$, for each pocket p_i the model generates m molecules $x_{i1}, x_{i2}...x_{ij}..., x_{im}$. Using docking software mentioned in Section 3.1, we calculate the affinity score between a molecule and a pocket: $S(x_{ij}, p_i)$. We define the binding ability of small molecules generated by the model for target p_i to target p_k as:

$$\text{BindingAbility}_{ik} = \mathbb{E}_{j\in(1,m)}[S(x_{ij}, p_i)] \tag{1}$$

In order to strengthen specificity and reduce the effect of PAINS fragments, we defined a novel metric, the delta score for SBDD. The metric is defined as follow:

$$\begin{aligned}\text{DeltaScore}(p_i) &= \text{BindingAbility}_{ii} - \text{BindingAbility}_{ik,k\neq i} \\ &= \mathbb{E}_{j\in(1,m)}[S(x_{ij}, p_i)] - \mathbb{E}_{j\in(1,m)}[S(x_{ij}, p_k)_{k\neq i}]\end{aligned} \tag{2}$$

We have also proposed a sampling technique in Appendix A, to improve the computational efficiency of approximating this expectation.

4.2 Experimental Results w.r.t. Delta Score

Table 2: Results on CrossDocked dataset with delta score. Best results are underlined.

Dataset	CrossDocked 2020	
Methods	**mean of mean (↓)**	**median of mean (↓)**
Ground Truth	<u>-0.810</u>	<u>-1.062</u>
AR	-0.535	-0.296
Pocket2mol	-0.309	-0.231
Shape2mol	0.052	-0.072
Targetdiff	-0.382	-0.505
Random_Zinc	-0.019	-0.018

We have conducted a reevaluation of state-of-the-art deep generative models on CrossDocked2020 with a focus on delta score. As indicated in Table 2, we have observed the correction of all exceptional results in Table 1. Firstly, there are significant differences among different methods, with ground truth being the best and random being the worst. This is reasonable because Random_Zinc could be considered as an unconditional random sampling method, its target aware binding performance is expected to close to zero. This reaffirms our belief that the delta score can incisively evaluate the conditional components of molecules, pinpointing those elements that genuinely and effectively engage with the target structure. Secondly, even the best generative models, e.g. AR and Targetdiff, shows significant differences from the ground truth, indicating that there is still a considerable research and development space in this direction.

5 Conclusion

While contemporary deep generative techniques have elevated the docking score, our analysis indicates that this enhancement predominantly pertains to the non-conditional aspects. Such improvements, we deduce, can be attributed to the methods learning biases that might mislead docking software, occasionally leading to the generation of anomalous molecular structures. We envision this pioneering metric—delta score—as a valuable addition to the current set of benchmarks. By providing deeper insights, we hope it will pave the way for more informed advancements in the realm of Structure-Based Drug Design.

References

[1] Leonardo G Ferreira, Ricardo N Dos Santos, Glaucius Oliva, and Adriano D Andricopulo. Molecular docking and structure-based drug design strategies. *Molecules*, 20(7):13384–13421, 2015.

[2] 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.

[3] Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks, 2022.

[4] 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. *arXiv preprint arXiv:2303.03543*, 2023.

[5] 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. 2023.

[6] Thomas A. Halgren, Robert B. Murphy, Richard A. Friesner, Hege S. Beard, Leah L. Frye, W. Thomas Pollard, and Jay L. Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of Medicinal Chemistry*, 47(7): 1750–1759, 2004.

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

[8] John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.

[9] A Lin, CJ Giuliano, A Palladino, KM John, C Abramowicz, ML Yuan, EL Sausville, DA Lukow, L Liu, AR Chait, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. sci. transl. med. 11: eaaw8412, 2019.

[10] Meng Liu, Youzhi Luo, Kanji Uchino, Koji Maruhashi, and Shuiwang Ji. Generating 3d molecules for target protein binding. *arXiv preprint arXiv:2204.09410*, 2022.

[11] Siyu Long, Yi Zhou, Xinyu Dai, and Hao Zhou. Zero-shot 3d drug design by sketching and generating. *Advances in Neural Information Processing Systems*, 35:23894–23907, 2022.

[12] 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.

[13] 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.

[14] Jiankun Lyu, John J Irwin, and Brian K Shoichet. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*, pages 1–7, 2023.

[15] Tomohide Masuda, Matthew Ragoza, and David Ryan Koes. Generating 3d molecular structures conditional on a receptor binding site with deep generative models, 2020.

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

[17] 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.

[18] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks, 2022.

[19] Trott, Oleg, Olson, Arthur, and J. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31 (2):NA–NA, 2009.

[20] Marcel L Verdonk, Jason C Cole, Michael J Hartshorn, Christopher W Murray, and Richard D Taylor. Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 2003.

[21] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

[22] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 2016.

[23] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

A Delta Score

To mitigate off-target effects, our new metric is designed to emphasize molecules with a specific binding affinity to the target protein pockets, as opposed to binding to multiple similar pockets without selectivity. As a result, we anticipate that all generated molecules should exhibit a negative score, as follows:

$$\begin{aligned}\text{BindingAbility}_{ii} - min_{k\in(1,n),k\neq i}\text{BindingAbility}_{ik} = \\ \mathbb{E}_{j\in(1,m)}[S(x_{ij},p_i)] - min_{k\in(1,n),k\neq i}\mathbb{E}_{j\in(1,m)}[S(x_{ij},p_k)] < 0\end{aligned} \tag{3}$$

Notice that to calculate $min_{k\in(1,n),k\neq i}\mathbb{E}_{j\in(1,m)}[S(x_{ij},p_k)]$, we need to perform docking for molecules with all possible pockets which is of quadratic complexity. To address this challenge, instead of docking molecules with every pockets, we randomly sample $\tilde{n}$ pockets outside of the pocket p_i: $\{p_{1'}...p_{\tilde{n}'}\} \subset \{p_1, p_2, ..., p_n\} \setminus \{p_i\}$ and define Delta Score for p_i as:

$$\text{DeltaScore}_{\tilde{n}}(p_i) = \mathbb{E}_{j\in(1,m)}[S(x_{ij},p_i)] - min_{k\in(1',\tilde{n}')}\mathbb{E}_{j\in(1,m)}[S(x_{ij},p_k)] \tag{4}$$

When $\tilde{n} = n-1$, the Delta Score is defined as equivalent to Eq. 2. In order to save computing resources, we set $\tilde{n} = 1$ in which case the delta score becomes:

$$\text{DeltaScore}_1(p_i) = \mathbb{E}_{j\in(1,m)}[S(x_{ij},p_i)] - \mathbb{E}_{j\in(1,m)}[S(x_{ij},p_k)_{k\neq i}] \tag{5}$$

It can be readily demonstrated that, in terms of statistical significance, this is equivalent to the difference between the model's binding affinity for a specific target and the model's average binding affinity for other targets:

$$\mathbb{E}_i[\text{DeltaScore}_1(p_i)] \equiv \mathbb{E}_i[\text{BindingAbility}_{ii} - \frac{1}{n-1}\sum_{k\in(1,n),k\neq i}\text{BindingAbility}_{ik}] \tag{6}$$

Proof. According to equation 2,

$$\mathbb{E}_{i\in(1,n)}[\text{DeletaScore}_1(p_i)] \tag{7}$$

$$= \mathbb{E}_{i\in(1,n)}\mathbb{E}_{j\in(1,m)}[S(x_{ij},p_i)] - \mathbb{E}_{i\in(1,n)}\mathbb{E}_{j\in(1,m)}[S(x_{ij},p_k)_{k\neq i}] \tag{8}$$

$$= \mathbb{E}_i[\text{BindingAbility}_{ii}] - \frac{1}{n}\sum_{i,k}^{i\neq k}\mathbb{E}_j[S(x_{ij},p_k)] \tag{9}$$

$$\equiv \mathbb{E}_i[\text{BindingAbility}_{ii}] - \frac{1}{n}\sum_{i}\frac{1}{n-1}\sum_{k}^{k\neq i}\mathbb{E}_j[S(x_{ij},p_k)] \tag{10}$$

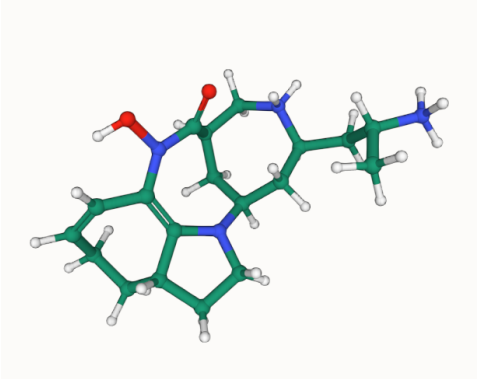
$$= \mathbb{E}_i[\text{BindingAbility}_{ii}] - \mathbb{E}_i[\frac{1}{n-1}\sum_{k\in(1,n),k\neq i}\text{BindingAbility}_{ik}] \tag{11}$$

$$= \mathbb{E}_i[\text{BindingAbility}_{ii} - \frac{1}{n-1}\sum_{k\in(1,n),k\neq i}\text{BindingAbility}_{ik}] \tag{12}$$

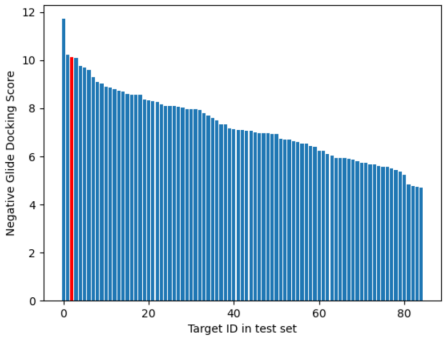
□

B Case Study

We choose one generated molecule against 4lfu_A_rec.pdb pocket for a more detailed display. It is questionable and has many issues as depicted in Figure 2. On the one hand, the molecular skeleton of the compound is complicated, consisting of multiple non-aromatic ring structures. The chemical reactions that could directly obtain this molecular skeleton are limited. The lipid rings exhibit poor water solubility and is easy to be metabolized. The N-hydroxamide is not common in drugs. On the other hand, we observe that it exhibits fairly good docking scores for the majority of the pockets in the test set, with an average of -7.02. Two pockets even surpass its actual target receptor on docking score, which indicates this molecule is highly likely to be a pan-assay interference compound or prone to off-target effect. Similar phenomenon occurs on almost all generated molecules with good docking scores, which indicates that only using docking score metric is far from enough to evaluate the effectiveness of current generative model.



(a) 2D structural diagram of the molecule



(b) Docking Scores against targets in the test set (multiplied by -1)

Figure 2: One molecule generated by Targetdiff Model against 4lfu_A_rec.pdb pocket. The left image shows its 2D structure. The right image shows its docking scores against pockets in the test set (after sorted). The column highlighted in red indicates its actual target.