PoseCheck: Generative Models for 3D Structure-based Drug Design Produce Unrealistic Poses

Anonymous Author(s)
Affiliation
Address
email

Abstract

Deep generative models for structure-based drug design (SBDD), where molecule generation is conditioned on a 3D protein pocket, have received considerable interest in recent years. These methods offer the promise of higher-quality molecule generation by explicitly modelling the 3D interaction between a potential drug and a protein receptor. However, previous work has primarily focused on the quality of the generated molecules themselves, with limited evaluation of the 3D *poses* that these methods produce, with most work simply discarding the generated pose and only reporting a "corrected" pose after redocking with traditional methods. Little is known about whether generated molecules satisfy known physical constraints for binding and the extent to which redocking alters the generated interactions. We introduce POSECHECK, an extensive analysis of multiple state-of-the-art methods and find that generated molecules have significantly more physical violations and fewer key interactions compared to baselines, calling into question the implicit assumption that providing rich 3D structure information improves molecule complementarity. We make recommendations for future research tackling identified failure modes and hope our benchmark will serve as a springboard for future SBDD generative modelling work to have a real-world impact. Our evaluation suite is easy to use in future 3D SBDD work and is available at `https://anonymous.4open.science/r/posecheck-358E`.

1 Introduction

Structure-based drug design (SBDD) [1, 2, 3] is a cornerstone of drug discovery. It uses the 3D structures of target proteins as a guide to designing small molecule therapeutics. The intricate atomic interactions between proteins and their ligands shed light on the molecular motifs influencing binding affinity, selectivity, and drug-like properties. Employing computational methods such as molecular docking [4, 5], molecular dynamics simulations [6], and free energy calculations [7], SBDD aids in the identification and optimization of potential drug candidates.

Deep generative models for SBDD have recently attracted considerable attention in the ML community [8, 9]. These models learn from vast compound databases to generate novel chemical structures with drug-like properties [10]. By explicitly integrating protein structure information, these models aim to generate ligands with a higher likelihood of binding to the target protein. In particular, advancements in geometric deep learning [11, 12, 13] have led to a new suite of generative methods, enabling the design of 3D molecules directly within the confines of the target protein [14, 15, 16, 17, 18]. These methods, which concurrently generate a molecular graph and 3D coordinates, provide the significant advantage of obviating the need for determining the 3D pose *post hoc* through traditionally slow molecular docking programs – at least in theory.

Submitted to the NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023). Do not distribute.

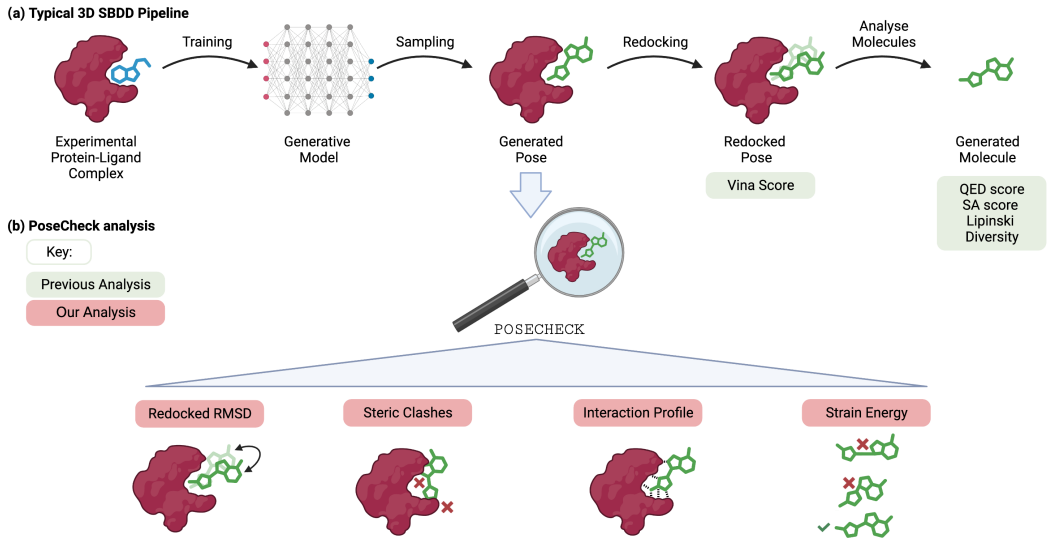

Figure 1: **Top:** Overview of a conventional pipeline of SBDD with 3D generative modelling. A generative model is usually trained using experimental or synthetic protein-ligand complexes, from which new molecules and poses can be sampled *de novo*. Typically, generated poses are discarded and redocked into the receptor, and primarily evaluated on 2D molecular graphs (e.g. QED). The effect of redocking on the final complex is often unknown, preventing understanding of the common failure modes of the trained model and therefore inhibiting progress. **Bottom:** the POSECHECK benchmarks for generated poses include pipeline-wide as well as component-wise metrics, enabling a targeted evaluation of each model component guiding further model development.

Assessing the quality of molecules generated by these methodologies is not straightforward, with little work on experimental validation, especially for *de novo* design [19]. Typical evaluation metrics (Figure 1a) focus primarily on the 2D graph of the generated molecules themselves, measuring their physicochemical properties (e.g. QED [20]) and adherence to drug discovery heuristics (e.g. Lipinski's Rule of Five [21]). For effective SBDD, we argue that it's equally essential to assess the quality of the generated *binding poses* and their capacity to satisfy known biophysical prerequisites for binding (Figure 1b). This perspective is essential if these methods are to serve as practical alternatives to traditional virtual screening approaches in SBDD.

We hypothesise that multiple failure modes, undetected by currently applied metrics, are inherent within these methods. The situation is further complicated by the common practice of disregarding the initially generated pose and then redocking the molecule to attain a potentially enhanced pose and only reporting these scores [16, 15, 18, 22]. This strategy tends to focus on presenting only the outcomes of the redocked molecule, and it is not clear whether molecules shown in figures are generated or redocked, making the accurate assessment of pose quality an increasingly intricate challenge.

Our primary contributions are summarized as follows: We introduce POSECHECK, a set of new biophysical benchmarks for SBDD models, expanding the traditional 'pipeline-wide' framework by integrating 'component-wise' metrics (i.e. generated and redocked poses), leading to comprehensive and precise model assessment. Utilizing this new framework, we evaluate a selection of high-performing machine learning SBDD methods, revealing two key findings: (1) generated molecules and poses often contain nonphysical features such as steric clashes, hydrogen placement issues, and high strain energies, and (2) redocking masks many of these failure modes. Based on these evaluations, we propose targeted recommendations to rectify the identified shortcomings. Our work thus provides a roadmap for addressing critical issues in SBDD generative modelling, informing future research efforts.

2 Background

Deep Generative Models for 3D Structure-based Drug Design Many works have recently tried to recast the SBDD problem as learning the 3D conditional probability of generating molecules given a receptor, allowing users to sample new molecules completely *de novo* inside a pocket. Common methods utilize Variational AutoEncoders (VAEs) [23], Generative Adversarial Networks (GANs) [24], Autoregressive (AR) models and recently Denoising Diffusion Probabilistic Models (DDPMs) [25]. LiGAN [14] uses a 3D convolutional neural network combined with a VAE model and GAN-style training. 3DSBDD [15] introduced an autoregressive (AR) model that iteratively samples from an atom probability field (parameterised by a Graph Neural Network) to construct a whole molecule, with an auxiliary network deciding when to terminate generation. Pocket2Mol [16] extended this work with a more efficient sampling algorithm and better encoder. DiffSBDD [17], DiffBP [26] and TargetDiff [17] are all conditional DDPMs conditioned on the 3D target structure. DecompDiff [27] is another diffusion model that decomposes the ligand into fragments for which it considers separate priors for the diffusion process. FLAG [22] chooses a fragment from a motif vocabulary based on the protein structure and composes it with other motifs into a final ligand in an iterative fashion. GraphBP [28] utilises an autoregressive flow model to formulate the ligand design as a sequential generation task.

Related work Guan et al. [17] perform limited analysis of small chemical sub-features, such as agreement to experimental atom-atom distances and the correctness of aromatic rings within the generated molecule. Baillif et al. [19] emphasize the necessity of 3D benchmarks for 3D generative models. However, both of these works study the molecules in isolation rather than the protein-ligand context. Both DecompDiff [27] and DiffBP [26] take steric clashes into account via their loss functions, but do not include steric clashes as a metric in their evaluation. TargetDiff [17] includes an analysis of Vina Scores but does not report any standard deviations on these. However, these standard deviations are critical in evaluating the performance of these models as we demonstrate in this paper.

The concurrent work PoseBusters [29] also focuses on benchmarking the biophysical plausibility of protein-ligand poses but focuses on evaluating *docking tools* instead of molecular generation models. They also find generalisation to new sequences to be poor.

3 Methods

In order to evaluate the quality of generated poses and their capacity to facilitate high-affinity protein-ligand interactions, we present a variety of computational methods and benchmarks in this section. These methodologies provide a thorough perspective on the poses produced and illuminate the ability of generative models to generate trustworthy and significant ligand conformations. Full implementation details are given in Appendix A.

Interaction fingerprinting Interaction fingerprinting is a computational method utilized in SBDD to represent and analyze the interactions between a ligand and its target protein. This approach encodes specific molecular interactions, such as hydrogen bonding and hydrophobic contacts, in a compact and easily comparable format – typically as a bit vector, known as a *interaction fingerprint* [30, 31]. Each element in the interaction fingerprint corresponds to a particular type of interaction between the ligand and a specific residue within the protein binding pocket. We compute interactions using the ProLIF library [30].

Steric clashes In the context of molecular interactions, the term *steric clash* is used when two neutral atoms come into closer proximity than the combined extent of their van der Waals radii [32]. This event indicates an energetically unfavourable [33], and thus physically implausible, interaction. The presence of such a clash often points towards the current conformation of the ligand within the protein being less than optimal, suggesting possible inadequacies in the pose design or a fundamental incompatibility in the overall molecular topology. Hence, the total number of clashes serves as a vital performance metric in the realm of SBDD. We stipulate a clash to occur when the pairwise distance between a protein and ligand atom falls below the sum of their van der Waals radii, allowing a clash tolerance of 0.5 Å.

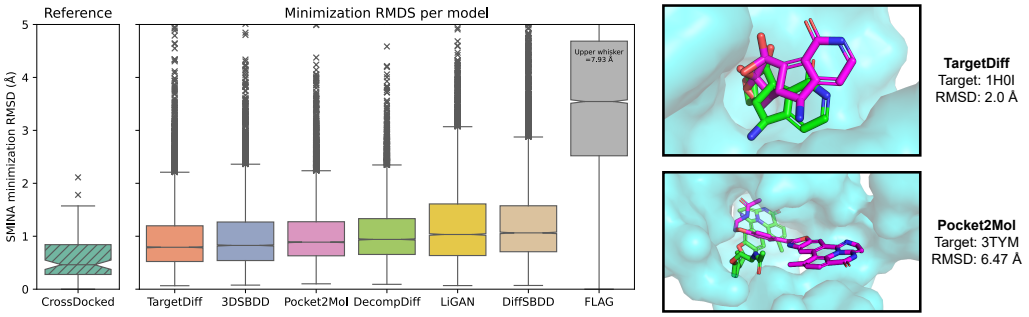

Figure 2: **Left:** RSMD between the generated and SMINA minimized poses for CrossDocked and all generative methods (note FLAG upper whister value is not shown to preserve a meaningful scale). **Right:** Examples of large conformational rearrangments in the ligand upon redocking.

Strain-energy Strain energy refers to the internal energy stored within a ligand as a result of conformational changes upon binding. When a ligand binds to a protein, both the ligand and the protein may undergo conformational adjustments to accommodate each other, leading to changes in their bond lengths, bond angles, and torsional angles. These changes can cause strain within the molecules, which can affect the overall binding affinity and stability of the protein-ligand complex [34]. Whilst there is always a trade-off between enthalpy and entropy, generally speaking, lower strain energy results in more favourable binding interactions and potentially more effective therapeutics. We calculate the strain energy as the difference between the internal energy of a relaxed pose and the generated pose (without pocket). Both relaxation and energy evaluation are computed using the Universal Force Field (UFF) [35] using RDKit.

Docking Our final assessment involves measuring the level of agreement between the docking programs and the molecules produced by the learned distribution in the generative model. Although this is the most coarse-grained approach we employ and docking programs come with their inherent limitations, they nevertheless contain useful proxies and serve as valuable tools for comparison. In this procedure, we redock the generated pose using SMINA [36]. Next, we compute the Root Mean Squared Deviation (RMSD) between the generated pose and the docking-predicted one across all generated molecules, thereby obtaining a distribution of RMSD values.

4 Results

4.1 Experimental Setup

In our study, we evaluate the quality of poses from seven recent methods: LiGAN [14], 3DSBDD [15], Pocket2Mol [16], TargetDiff [17], DiffSBDD [18], DecompDiff [27] and FLAG [22]. All models were trained on the CrossDocked2020 [37] dataset using the dataset splits computed in Peng et al. [16], which used a train/test split of 30% sequence identity to give a test set of 100 target protein-ligand complexes which we use for evaluation. For each model, we sampled 100 molecules per target. We give a more detailed overview of the CrossDocked dataset and its limitations in Appendix A.

4.2 Agreement with docking scoring functions

Results To discern whether the generated poses/binding modes produced by these models correspond to overall low energy states with few physical violations, our preliminary analysis involves determining the extent to which minimized poses preserve information from the initially generated binding mode. Therefore, we proceed to compute the RMSD between the model-generated pose and SMINA-minimized pose [36], with a lower RMSD value denoting a higher degree of agreement.[1]

The distributions of SMINA-minimization RMSDs of various methods are illustrated in Figure 2. We first consider CrossDocked as a baseline, which has a mean minimization RMSD of 0.59 Å.

[1]To provide perspective, it's worth noting that a carbon-carbon bond generally measures 1.54 Å in length.

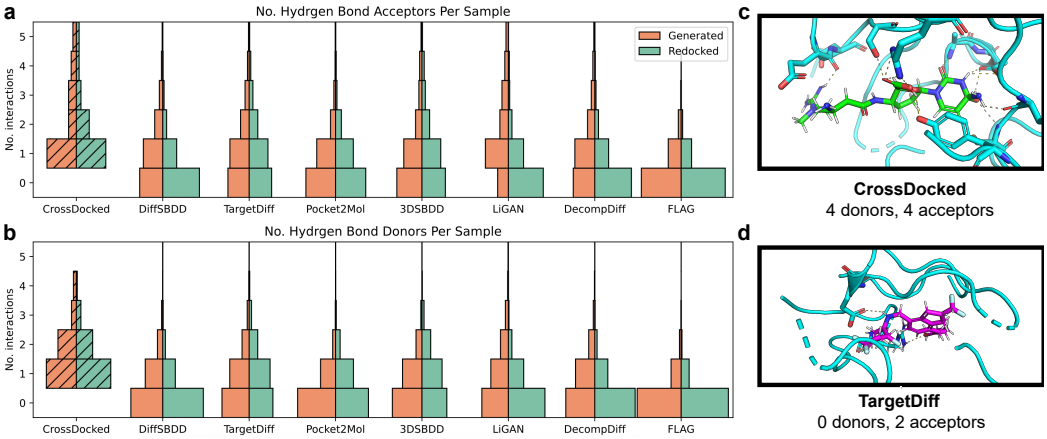

Figure 3: Interactions between protein and ligands as seen in generated poses (orange) and redocked poses (green). The frequency of (a) hydrogen bond acceptors and (b) hydrogen bond donors are considered. We find that generative models have significant trouble making hydrogen bond interactions compared to baseline (shaded boxes). Vertical histogram box sizes are normalised along the x-axis such that all have the same area. (c) Example from CrossDocked with large hydrogen bonding network. (d) Typical example from a generative models with low number of HBs.

Given that all the generative models were trained on these poses, we would expect to observe similar performance. However, we find that all methods (except FLAG) have a mean score between 0.94 and 1.28 Å, suggesting that the generated binding poses are very far from low-energy states. We observe little correlation between method types here except for the two similar AR models, 3DSBDD and Pocket2Mol, which obtain mean RMSDs of 0.99 and 1.02 Åresspectively. FLAG is the most egregious example with on average 3.64 ÅRMSD during minimization and a maximum value of 10.72 Å, an extreme value for local minimisation.

Discussion These findings raise concerns for several reasons. They expose the minimal concordance between the binding models learned by these methods and the established SMINA methodology [5], despite it being the source of training data. More critically, they underline the lack of accurate evaluations of generative models' capability to produce realistic binding poses; instead, these models tend to generate drug-like molecules with vague binding modes, later rectified through docking.

We also calculated the RMSD between the generated and highest affinity redocked pose but were not able to discern any reasonable signal-to-noise over the baseline dataset. We hypothesise that this may be due to the fact that Francoeur et al. [37] provided up to 20 poses for every ligand, resulting in 22.5 million complexes, and the processing done in Peng et al. [16] is not clear on which poses they chose, meaning these models may not have been trained on the lowest affinity poses.

4.3 Protein-ligand interaction analysis

Evaluation Below describe the classes of interaction that we evaluate. **Hydrogen bonds** (HBs) are a type of interaction that occurs between a hydrogen atom that is bonded to a highly electronegative atom, such as nitrogen, oxygen, or fluorine [38]. They are key to many protein-ligand interactions [39] and require very specific geometries to be formed [40]. The directionality of HBs confers unique identities upon the participating atoms: hydrogen atoms attached to electronegative elements are deemed 'donors', whilst the atom accepting the HB is termed an 'acceptor'. **Van der Waals contacts** (vdWs) are interactions that occur between atoms that are not bonded to each other. These forces can be attractive or repulsive and are typically quite weak [41]. However, they can be significant when many atoms are involved, as is typical in protein-ligand binding [42]. **Hydrophobic interactions** are non-covalent interactions that occur between non-polar molecules or parts of molecules in a water-based environment. They are driven by the tendency of water molecules to form hydrogen bonds with each other, which leads to the exclusion of non-polar substances. This exclusion principle prompts these non-polar regions to orient away from the aqueous environment and towards each other [43], thereby facilitating the association between protein and ligand molecules [44].

Results Distributions of hydrogen bonding interactions are shown in Figure 3. We consider whether our generative models can design molecules with adequate hydrogen bonding and find that no method can match or exceed the baseline. In the reference set, CrossDocked, the modal number of HBs for both acceptors and donors is 1, with means of 2.23 and 1.66 for acceptors and donors respectively. Strikingly, we find that in all generated poses for all models (except LiGAN HB acceptors) the *most common number of HB acceptors and donors is 0*, with means varying between 0.36-1.73 for HB acceptors and 0.26-0.85 for HB donors. We find an average difference of 0.50 and 0.81 HBs between the best-performing models and the baseline for acceptors and donors respectively. Results for Van der Waals contacts and hydrophobic interactions are closer to the dataset baseline (see Appendix Figure 6), possibly as these are easier to form.

Discussion Conventional wisdom would suggest that many minor imperfections in the generated pose would be simply fixed by redocking the molecule (e.g. moving an oxygen atom slightly to complete a hydrogen bond.) We find this is in fact rarely the case, with redocking sometimes being significantly deleterious (see examples of LiGAN in Figure 3), suggesting that there are either limitations in the docking function used or, more likely, the generated interaction was physically implausible to begin with.

4.4 Clash scores

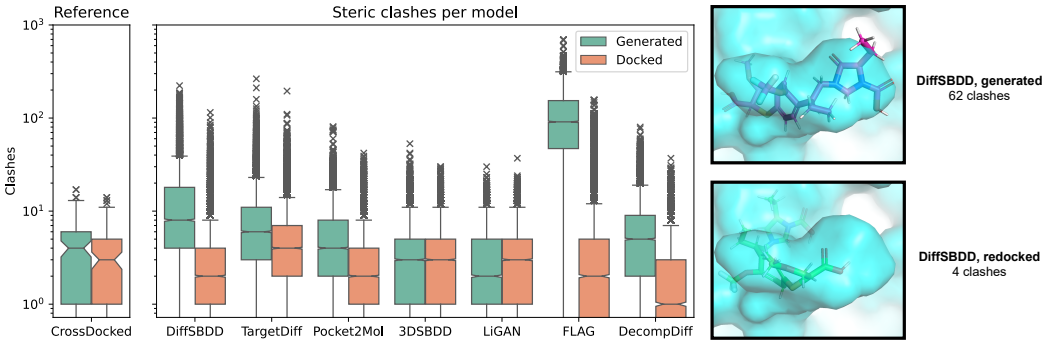

Figure 4: **Left:** number of steric clashes for the CrossDocked reference dataset as well as for the molecules generated by each model, both before and after redocking. **Right:** examples of a generated pose (magenta) and the same pose after redocking (green).

Results Figure 4 presents the results of the steric clash analysis. In summary, the latest methods, particularly those employing diffusion models and fragment libaries, exhibit poor performance in terms of steric clashes compared to the baseline, with a significant number of outliers. Although redocking mitigates clashes to some extent, it does not always resolve the most severe cases.

The CrossDocked test set has a low number of clashes with few extreme examples, with a mean of 4.59, upper quantile of 6 and maximum value of 17. In terms of generated poses, the older methods perform best, with 3DSBDD and LiGAN having means of 3.79 and 3.40 clashes respectively. Pocket2Mol, an extension of 3DSBDD, performs worse with a mean clash score of 5.62 and upper quantile of 8 clashes. Finally, the diffusion-based approaches perform poorly with mean clash scores of 15.33, 9.03 and 7.13 for DiffSBDD, TargetDiff and DecompDiff respectively. The tail end of their distributions is also high, with the methods having upper quantiles of 18, 11 and 9 clashes respectively, with TargetDiff having the worst case of 264 steric clashes. FLAG has the worst generated clash scores by far, with mean and median clash scores of 110.96 and 91 respectively. Redocking the molecules generally fixed many clashes and improved scores, especially for FLAG, where the mean clash score improves from 110.96 to 5.55. The mean clash score for Pocket2Mol improves from 5.62 to 2.98, TargetDiff from 9.08 to 5.79 and DiffSBDD from 15.34 to 3.61.

Discussion Interestingly, DiffSBDD and TargetDiff, which are considered state-of-the-art based on mean docking score evaluations [17, 18], exhibit subpar performance in their number of clashes. They aim to learn atom position distributions without explicit constraints on final placements. While DiffSBDD starts with a performance deficit, its enhanced clash mitigation during redocking elevates its results to match the baseline, highlighting methodological distinctions between it and TargetDiff.

Notably, 3DSBDD and LiGAN show low clash scores, with the former positioning atoms within a predefined voxel grid [15] and the latter applying a clash loss [14]. DecompDiff also applies a steric clash loss (but does not directly measure clashes in the corresponding publication) [27] and performs best out of all the diffusion-based approaches. Generated molecules for FLAG were most egregious here; we speculate this is a result of first choosing a fragment from a fragment vocabulary using a softmax function and then forcing the placement of the fragment [22], regardless of whether it fits sterically.

Our findings affirm the assumption that redocking alleviates many minor clashes, akin to the force-field relaxation step in AlphaFold2 [45]. We initially speculated that molecules with clashes exceeding 100 had been mistakenly generated inside the protein pocket. Yet, we often discovered fragments within highly constrained nooks, especially worsened with the addition of hydrogen atoms.

Limitations An important consideration to bear in mind is that proteins are not entirely rigid receptors. They can often experience limited conformational rearrangements to accommodate molecules of varying shapes and sizes [46]. Consequently, conducting generation and redocking in a rigid receptor environment may not yield accurate scores for potentially plausible molecules. Note all these results are with a *generous* clash tolerance of 0.5 Å (roughly half the vdW radii of a hydrogen atom), in order to be able to resolve differences between methods.

4.5 Strain energy

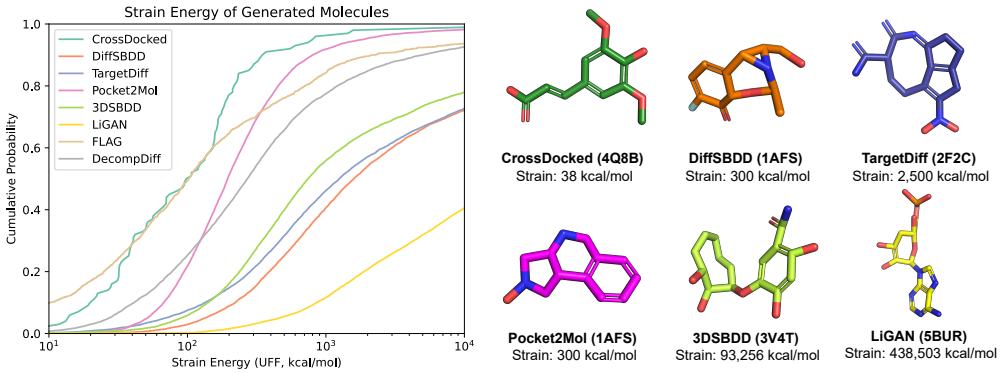

Figure 5: **Left:** CDF of strain energies. **Right:** Examples of molecules with high strain energy.

Results To conclude our study, we provide an analysis of the strain energy [34] of the generated poses. Force field relaxation before docking is a common post-processing step of many generative SBDD pipelines, masking some potential issues with the generated geometries less clear. This allows us to evaluate the generated molecules for undesirable properties like unrealistic bond distance or impossible geometries in rings.

Figure 5 displays the cumulative density function (CDF) of strain energy for the generated molecules, with the CrossDocked dataset serving as a baseline (Note: the x-axis is on a logarithmic scale). We focus on median values in our discussion since they are more representative in this context due to the presence of extreme outliers, with *mean* values ranging from approximately 10^4 to 10^{15} kcal/mol. None of the generative methods yields molecules exhibiting strain energy close to that of the test set, which has a median strain energy of 102.5 kcal/mol.

Discussion Intriguingly, both of the diffusion-based methodologies (DiffSBDD and TargetDiff) perform similarly poorly, reporting median values of 1243.1 and 1241.7 kcal/mol, respectively. This could suggest issues with the currently used noised schedules [47] of these methods for ultra-precise atom position refinement (discussed in Section C). 3DSBDD performs to the same order of magnitude, with a median strain energy of 592.2 kcal/mol, suggesting that placing atoms into a discretized voxel space [15], while good for avoiding clashes, has a detrimental impact on the strain energy.

FLAG performs the best by far here with a median of 101.1 kcal/mol. We believe this due to most of the bond angles and distances already consisting of idealised geometries when the fragments are

initialized for incorporation into the molecule. Out of the other methods, Pocket2Mol performs the best in terms of strain energy, with a median of 194.9 kcal/mol. The method provides perhaps the finest-grained control over exact coordinates generated, by first choosing a focal atom and then generating a new atom coordinate directly using an equivariant neural network [13, 16], which may allow for more precise placement. LiGAN exhibits the highest strain energy, with a median value of 18693.8 kcal/mol, indicating the poorest performance in this context.

Limitations The exceedingly high strain energy values observed in this scenario should be approached with considerable prudence. For comparison, the combustion of TNT releases approximately 815 kcal/mol. [48]. This data is not to be perceived as absolute, but rather illustrative of the extent to which our generated geometries deviate markedly from the standard distribution for the force field. This further underscores the existing issues. It is also conceivable that these poses might not even be initialized within more sophisticated, high-fidelity force fields [49].

5 Conclusion

In conclusion, this work presents a comprehensive exploration of structure-based drug design (SBDD) methodologies with deep generative models. We advocate for the need to consider *both* the quality of the generated molecules *and* the quality of the binding poses in these models, calling for an expanded evaluation of SBDD. The application of deep generative models in SBDD holds promise for developing innovative drug-like molecules. However, for SBDD approaches to realise that potential, we must establish a rigorous evaluation regimen of both the generated molecules and their interaction with the target – as proposed in this paper. Our research provides a solid evaluation regimen for future advancements in this field and we hope that it stimulates further development towards more efficient drug discovery processes.

References

[1] Tom L Blundell. Structure-based drug design. *Nature*, 384(6604 Suppl):23–26, 1996.

[2] 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.

[3] Amy C Anderson. The process of structure-based drug design. *Chemistry & biology*, 10(9): 787–797, 2003.

[4] Oleg Trott and Arthur J Olson. 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.

[5] Amr Alhossary, Stephanus Daniel Handoko, Yuguang Mu, and Chee-Keong Kwoh. Fast, accurate, and reliable molecular docking with quickvina 2. *Bioinformatics*, 31(13):2214–2216, 2015.

[6] John L Klepeis, Kresten Lindorff-Larsen, Ron O Dror, and David E Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Current opinion in structural biology*, 19(2):120–127, 2009.

[7] Christophe Chipot and Andrew Pohorille. *Free energy calculations*, volume 86. Springer, 2007.

[8] Yuanqi Du, Tianfan Fu, Jimeng Sun, and Shengchao Liu. Molgensurvey: A systematic survey in machine learning models for molecule design. *arXiv preprint arXiv:2203.14500*, 2022.

[9] Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023.

[10] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[11] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34 (4):18–42, 2017.

[12] Kenneth Atz, Francesca Grisoni, and Gisbert Schneider. Geometric deep learning on molecular representations. *Nature Machine Intelligence*, 3(12):1023–1032, 2021.

[13] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

[14] Tomohide Masuda, Matthew Ragoza, and David Ryan Koes. Generating 3d molecular structures conditional on a receptor binding site with deep generative models. *arXiv preprint arXiv:2010.14442*, 2020.

[15] 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.

[16] 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.

[17] 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.

[18] Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

[19] Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Deep generative models for 3d molecular structure. *Current Opinion in Structural Biology*, 80:102566, 2023.

[20] 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.

[21] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 64:4–17, 2012.

[22] Zaixi Zhang, Yaosen Min, Shuxin Zheng, and Qi Liu. Molecule generation for target protein binding with structural motifs. In *The Eleventh International Conference on Learning Representations*, 2022.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[26] Haitao Lin, Yufei Huang, Meng Liu, Xuanjing Li, Shuiwang Ji, and Stan Z Li. Diffbp: Generative diffusion of 3d molecules for target protein binding. *arXiv preprint arXiv:2211.11214*, 2022.

[27] 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.

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

[29] Martin Buttenschoen, Garrett M Morris, and Charlotte M Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *arXiv preprint arXiv:2308.05777*, 2023.

[30] Cédric Bouysset and Sébastien Fiorucci. Prolif: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13:1–9, 2021.

[31] Gilles Marcou and Didier Rognan. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of chemical information and modeling*, 47(1): 195–207, 2007.

[32] Srinivas Ramachandran, Pradeep Kota, Feng Ding, and Nikolay V Dokholyan. Automated minimization of steric clashes in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 79(1):261–270, 2011.

[33] Rosa Buonfiglio, Maurizio Recanatini, and Matteo Masetti. Protein flexibility in drug discovery: from theory to computation. *ChemMedChem*, 10(7):1141–1148, 2015.

[34] Emanuele Perola and Paul S Charifson. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *Journal of medicinal chemistry*, 47(10):2499–2510, 2004.

[35] Anthony K Rappé, Carla J Casewit, KS Colwell, William A Goddard III, and W Mason Skiff. Uff, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *Journal of the American chemical society*, 114(25):10024–10035, 1992.

[36] David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

[37] 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.

[38] George C Pimentel and AL McClellan. Hydrogen bonding. *Annual Review of Physical Chemistry*, 22(1):347–385, 1971.

[39] Deliang Chen, Numan Oezguen, Petri Urvil, Colin Ferguson, Sara M Dann, and Tor C Savidge. Regulation of protein-ligand binding affinity by hydrogen bond pairing. *Science advances*, 2(3): e1501240, 2016.

[40] ID Brown. On the geometry of o–h-o hydrogen bonds. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(1):24–31, 1976.

[41] Ylva Andersson, Erika Hult, Henrik Rydberg, Peter Apell, Bengt I Lundqvist, and David C Langreth. Van der waals interactions in density functional theory. *Electronic Density Functional Theory: Recent Progress and New Directions*, pages 243–260, 1998.

[42] Elizabeth Barratt, Richard J Bingham, Daniel J Warner, Charles A Laughton, Simon EV Phillips, and Steve W Homans. Van der waals interactions dominate ligand- protein association in a protein binding site occluded from solvent water. *Journal of the American Chemical Society*, 127(33):11827–11834, 2005.

[43] Emily E Meyer, Kenneth J Rosenberg, and Jacob Israelachvili. Recent progress in understanding hydrophobic interactions. *Proceedings of the National Academy of Sciences*, 103(43):15739–15746, 2006.

[44] Rohan Patil, Suranjana Das, Ashley Stanley, Lumbani Yadav, Akulapalli Sudhakar, and Ashok K Varma. Optimized hydrophobic interactions and hydrogen bonding at the target-ligand interface leads the pathways of drug-designing. *PloS one*, 5(8):e12029, 2010.

[45] 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.

[46] Andrew M Davis and Simon J Teague. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angewandte Chemie International Edition*, 38(6):736–749, 1999.

[47] Ting Chen. On the importance of noise scheduling for diffusion models. *arXiv preprint arXiv:2301.10972*, 2023.

[48] Wm H Rinkenbach. The heats of combustion and formation of aromatic nitro compounds. *Journal of the American Chemical Society*, 52(1):115–120, 1930.

[49] Bernard R Brooks, Charles L Brooks III, Alexander D Mackerell Jr, Lennart Nilsson, Robert J Petrella, Benoît Roux, Youngdo Won, Georgios Archontis, Christian Bartels, Stefan Boresch, et al. Charmm: the biomolecular simulation program. *Journal of computational chemistry*, 30 (10):1545–1614, 2009.

[50] Irina Kufareva, Andrey V Ilatovskiy, and Ruben Abagyan. Pocketome: an encyclopedia of small-molecule binding sites in 4d. *Nucleic acids research*, 40(D1):D535–D540, 2012.

[51] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[52] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

[53] John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *bioRxiv*, pages 2022–12, 2022.

[54] Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation. *arXiv preprint arXiv:2302.02277*, 2023.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] In the repository

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] Publicly available.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A CrossDocked Dataset

R2Q3

The CrossDocked dataset is a standard dataset used in the field of generative modelling for structure-based drug design [37]; since the models benchmarked here were trained on this dataset, it is the benchmarking dataset of choice. It was originally created by clustering PDB structures by "pocket similarity' via Pocketome [50], i.e. grouping structures with similar ligand binding sites together. To expand the dataset beyond this initial data, all ligands with a molecular weight < 1000 Da that were associated with a given pocked were docked into each receptor assigned to that pocket via the docking tool smina [36]. This cross-docking process results in the basis dataset CrossDocked 2020 [37], which contains 2,922 pockets, 18,450 complexes and 13,839 ligands, together comprising around 22.5 million poses (i.e. protein-ligand structures).

Most generative models are however not trained on this raw dataset, but on a filtered version of it, following the procedure of the Pocket2Mol model [16]. As a quality control, data points whose binding pose RMSD is greater than 1 were filtered out. This leads to a filtered dataset with 184,057 data points. The mmseq2 program [51] was used to cluster data at 30% identity, and training and test sets were created by randomly drawing 100,000 protein-ligand pairs for training and 100 proteins from the remaining clusters for testing.

The 100 proteins comprising the test set are on average around 320 residues long, with the biggest protein having a length of 752 residues.

B Extended Implementation

B.1 Methods Implementation

All generative methods accessed were trained using the same dataset and splits as proposed in Peng et al. [16]. Docking protocols were done using the SMINA settings decribed in the original CrossDocked paper [37].

B.2 Procedure of model reproduction

For generated poses, we sourced molecules from Schneuing et al. [18] for DiffSBDD, and Guan et al. [17] for CrossDocked, TargetDiff, Pocket2Mol, 3DSBDD and LiGAN (where they provide generated poses but we additionally perform our own redocking).

For FLAG [22], no weights were provided so we retrained the model as described in Zhang et al. [22] using the code and config file available at `github.com/zaixizhang/FLAG`. When sampling, we found that generation was attempted 100 times per target and then any molecules with fewer than 8 atoms were discarded. This ended up encompassing the majority of molecules, resulting in small test sizes, so we implemented a while loop to sample 100 molecules whilst keeping faithful to the filtering used in the codebase. Having modified the code to work on GPU, sampling 100 targets took about 1-2 minutes per target on a single A100 GPU.

For DecompDiff [27], we use the official implementation with the published weights available at `github.com/bytedance/DecompDiff`. We sampled 100 samples for each of the 100 targets using the `sample_diffusion_drift.py` script in `ref_prior` mode. With the provided code, sampling 100 targets took about 20-30 minutes per target on a single A100 GPU.

C Recommendations for future work

Exploring reduced-noise sampling strategies Interestingly, both diffusion-based works (DiffSBDD and TargetDiff) performed similarly in terms of strain energy (see Section 4.5). We hypothesize this may be due to the injection of random noise into the coordinate features at all but the last step of stochastic gradient Langevin dynamics samplings [52], making it challenging to construct precise bond angles etc. Here, inspiration could be taken from protein design. For example, Chroma develops a low-temperature sampling regime to reduce the effect of noise [53], FrameDiff effectively scales

down injected noise [54], both resulting in a substantial increase in sample quality with an acceptable decrease in sample diversity.

Heavily penalise steric clashes during training All evaluated methods frequently create steric clashes, resulting in physically unrealizable samples. We suggest that mitigating steric clashes is key for the next generation of SBDD models. This could be done via extra loss terms, for example, by including a distogram loss as in AlphaFold2 [45] or the steric clash loss in LiGAN [14] and DecompDiff [27] (note that later method does note explicitly measure clashes). A similar loss-based approach has been effective in mitigating chain-breaks diffusion models for protein backbone design [54].

Consider representing hydrogens Virtually all work in ML for structural biology chooses to not explicitly represent hydrogen atoms [45, 16, 15, 54, 18, 17], under the assumption that they can be *implicitly* learned and reasoned over with deep neural networks. However, our analysis of hydrogen bond networks within generated molecules found that generative methods struggle to handle the precise geometries required to make a hydrogen bond [40] (even when redocked). Despite the increased computational cost, we therefore recommend that future work explores their inclusion.

D Additional Figures

D.1 Interactions analysis

We include the comparisons between generative method against baselines for both Van der Waals contacts and hydrophobic interactions, both for generated redocked poses in Figure 6.

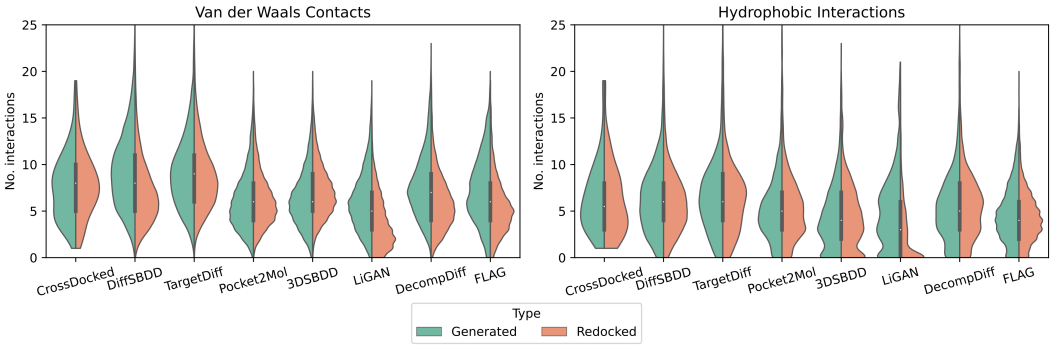

Figure 6: Extended analysis of the interaction profiles of the generated molecules for the different methods. While the focus in the main text was on hydrogen bonds, the results in this figure include Van der Waals Contacts and hydrophobic interactions, reported for both the generated as well as the redocked pose.

D.2 Redocking and clashes analysis

In Figure 7, we provide the per target redocking RMSDs per method. Figure 8 and 9 show the number of steric clashes per target for the generated and redocked poses respectively.

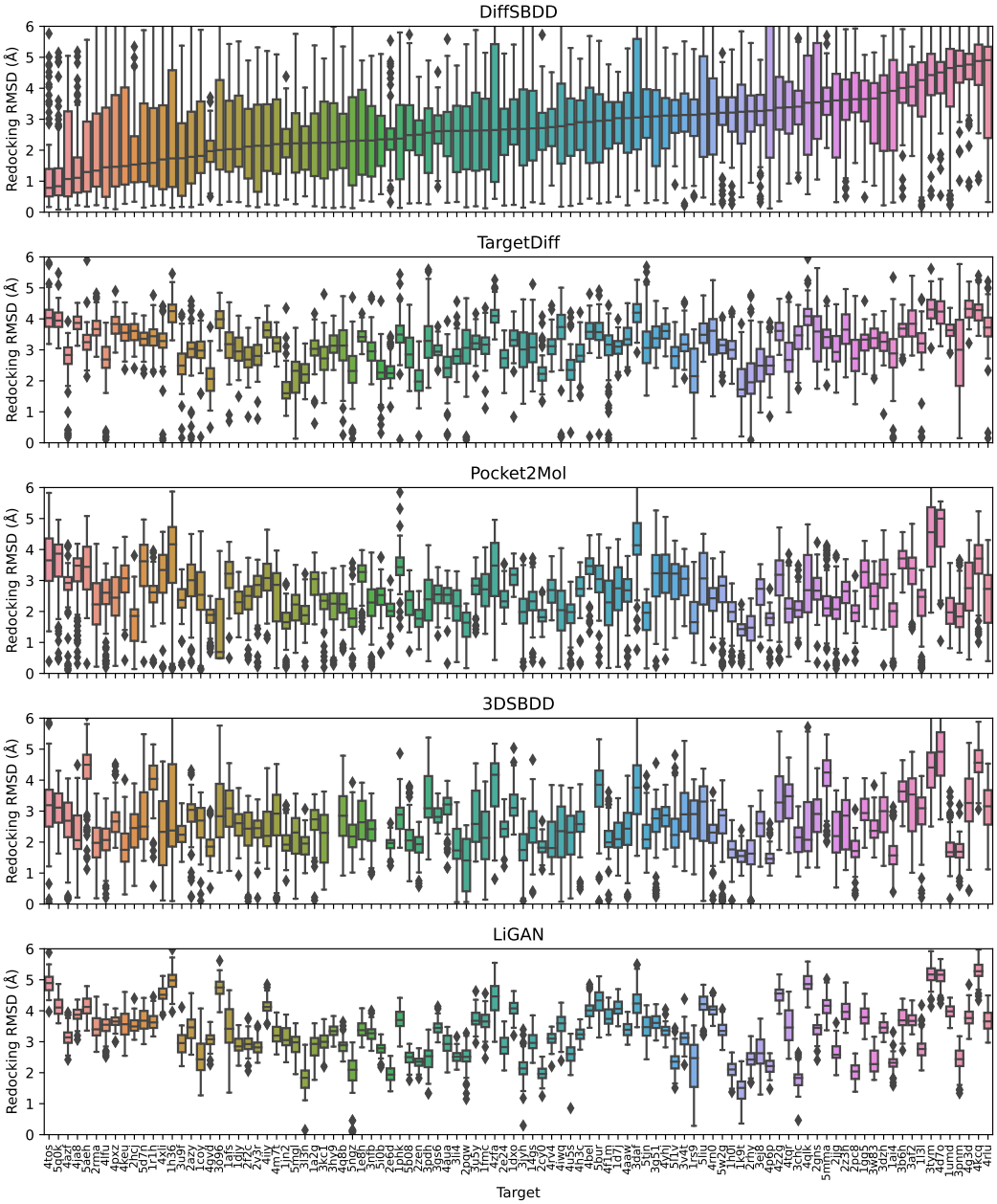

Figure 7: Redocking RMSD per method per target for CrossDocked test set. Order is determined arbitrarily by median score per target for DiffSBDD.

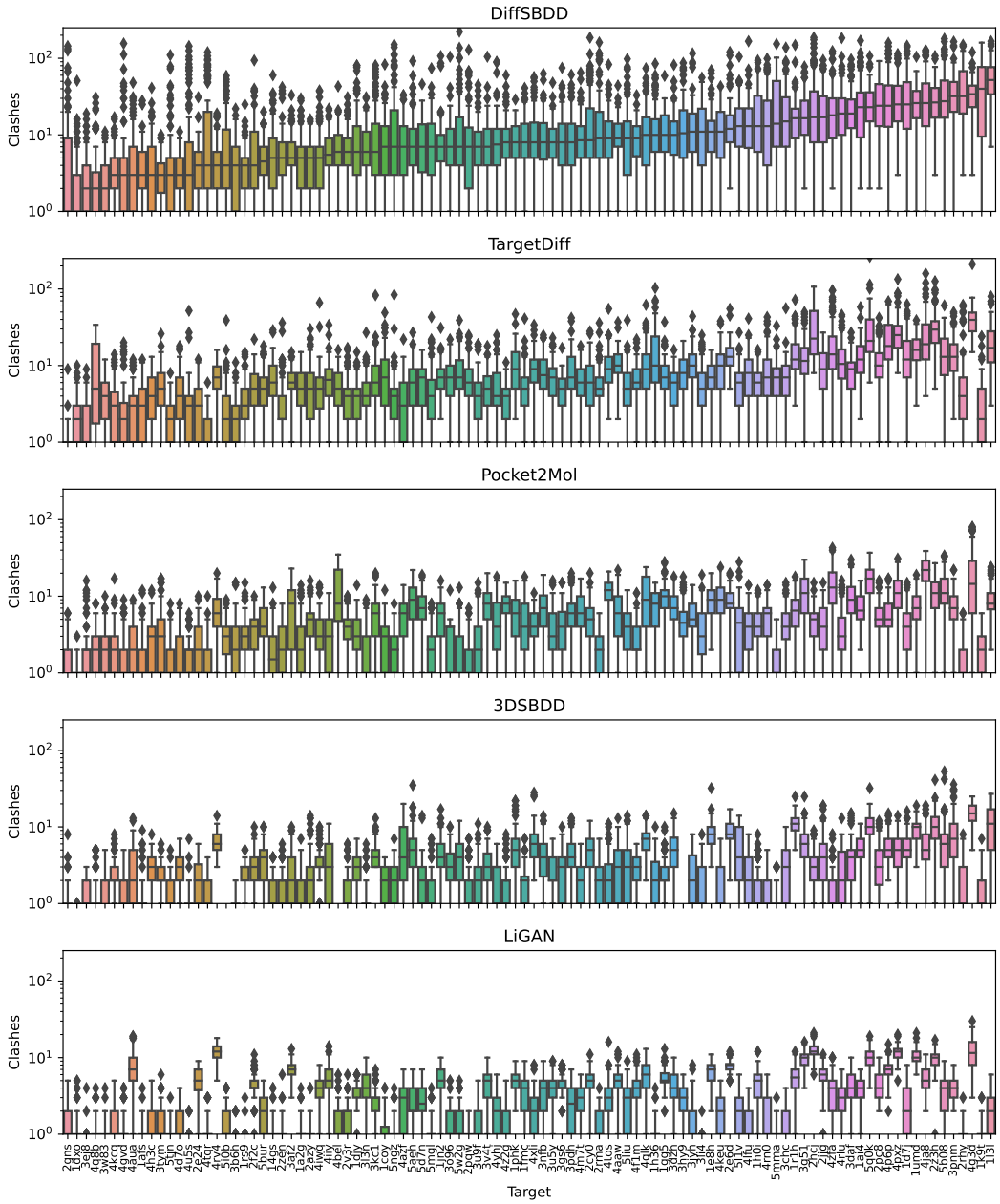

Figure 8: Steric clashes per method per target for generated poses in the CrossDocked test set. Order is determined arbitrarily by median score per target for DiffSBDD.

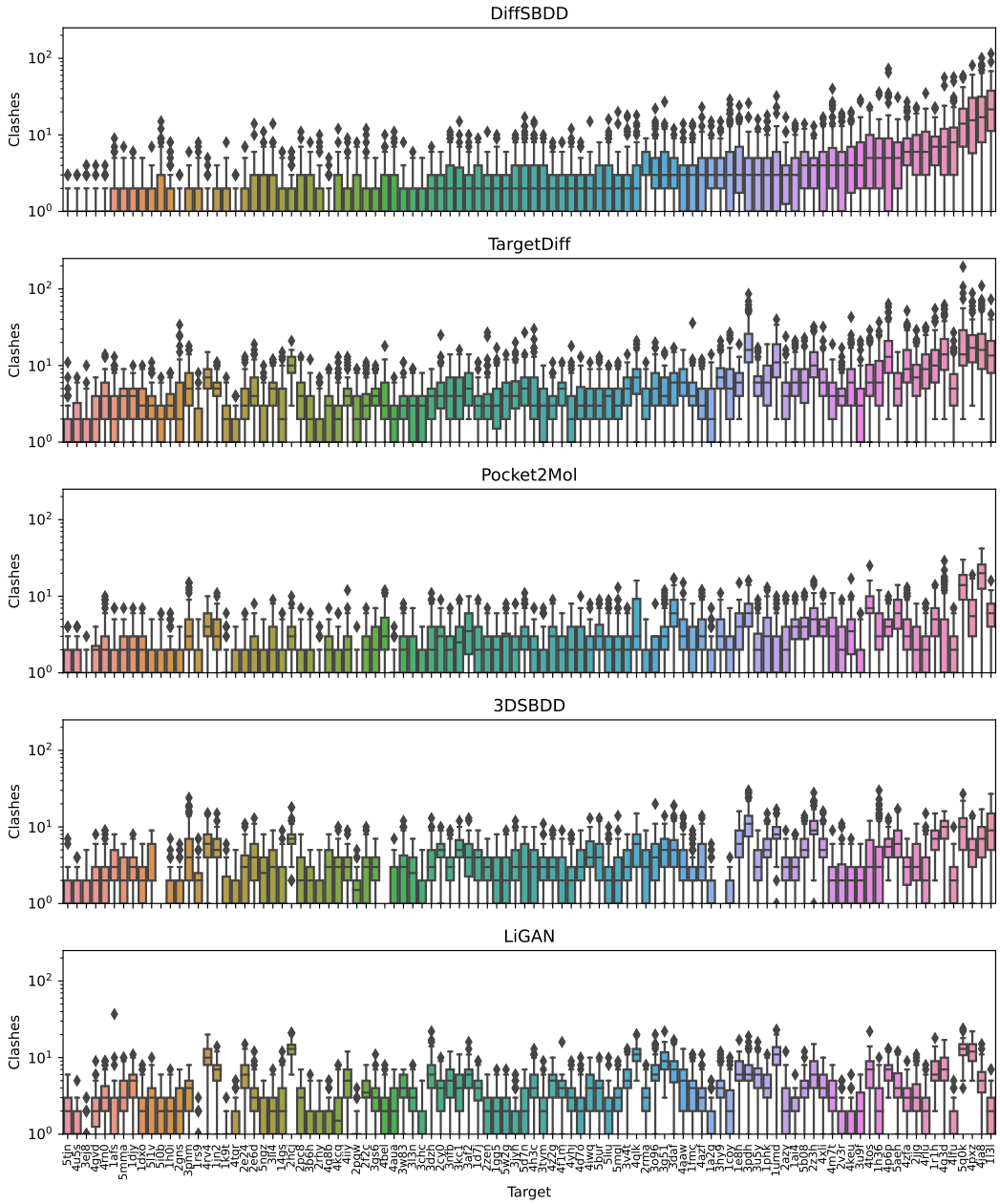

Figure 9: Steric clashes per method per target for redocked poses in the CrossDocked test set. Order is determined arbitrarily by median score per target for DiffSBDD.