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

Charles Harris*
University of Cambridge
cch57@cam.ac.uk

Kieran Didi
University of Cambridge
ked48@cam.ac.uk

Arian R. Jamasb
University of Cambridge
arj39@cam.ac.uk

Chaitanya K. Joshi
University of Cambridge
ckj24@cam.ac.uk

Simon V. Mathis
University of Cambridge
svm34@cam.ac.uk

Pietro Lio
University of Cambridge
pl219@cam.ac.uk

Tom L. Blundell
University of Cambridge
tlb20@cam.ac.uk

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 www.github.com/cch1999/posecheck.

1 Introduction

Structure-based drug design (SBDD) leverages knowledge of the 3D structure of a target protein to design highly potent and specific drug compounds [1, 2, 3]. Recent advancements in machine learning, particularly deep generative models [4, 5] and geometric deep learning [6], have led to an explosion of methods combining both to perform SBDD using 3D generative modelling [7, 8, 9, 10, 11].

Assessing the quality of molecules generated by these methodologies is not straightforward, with most metrics focusing on the 2D graph of the generated molecule themselves (e.g. QED [12]) and

*Correspondance to cch57@cam.ac.uk

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

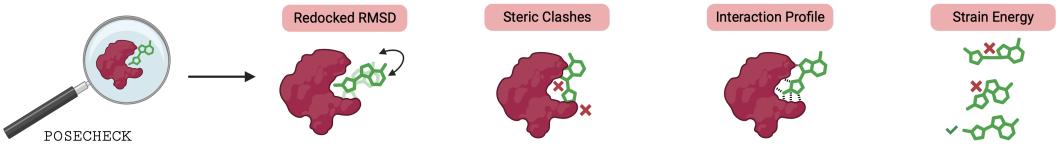

Figure 1: Overview of the POSECHECK pipeline.

many works using high mean docking score to claim state-of-the-art performance. For effective SBDD, we argue that it's equally important to assess the quality of the generated *binding poses* and their capacity to satisfy known biophysical prerequisites for binding. 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 pervasive within these methods.

We introduce POSECHECK, a set of new biophysical benchmarks for SBDD models. 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 Methods

We propose a suite of benchmarks to evaluate the quality of generated poses from 3D SBDD models.

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* [13, 14], allowing for easy comparaison between complexes.

Steric clashes A *steric clash* is when two atoms come into closer proximity than the sum of their atomic radii [15], which is highly unfavourable [16]. 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.

Strain-energy Strain energy refers to the internal energy stored within a ligand as a result of conformational changes upon binding. These changes can cause strain within the molecules, which can affect the overall binding affinity and stability of the protein-ligand complex [17].

Docking Finally, we measure the level of agreement between docking software and the molecules produced by the learned distribution in the generative model. Although this is the most coarse-grained

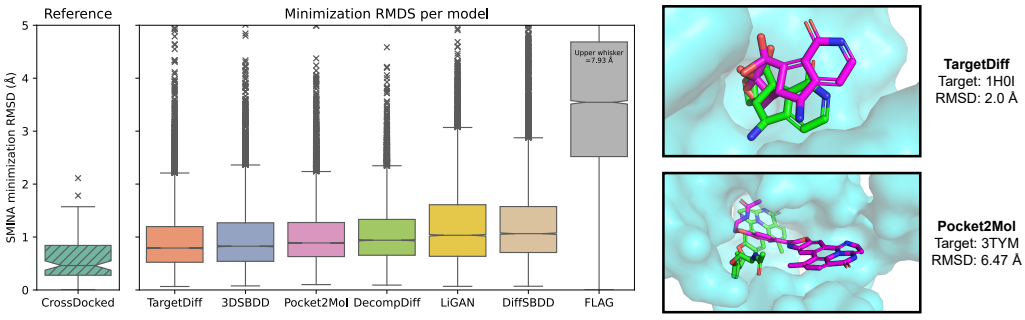

Figure 2: **Left:** RSMD between the generated and SMINA minimized poses for CrossDocked and all generative methods (note FLAG upper whister value is not shown to preserve a meaningful scale). **Right:** Examples of large conformational rearrangements in the ligand upon redocking. **None of the methods is able to generate poses in as low energy a state as the training data**

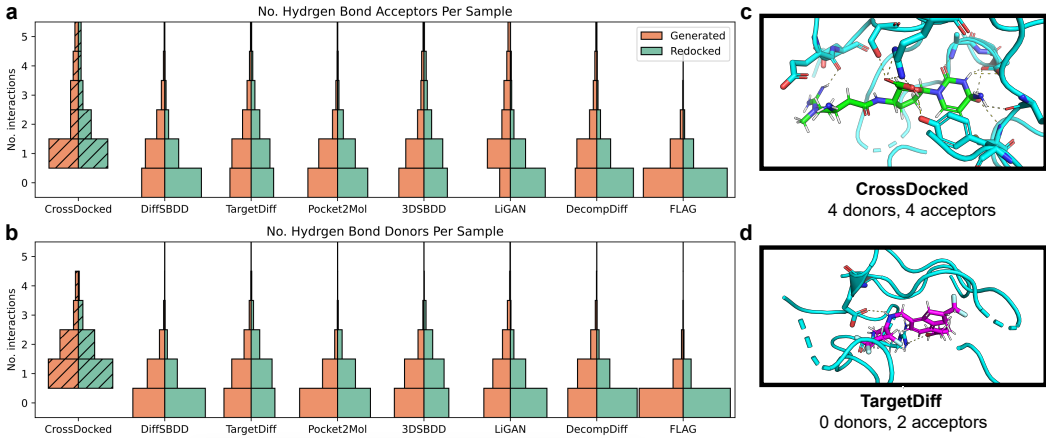

Figure 3: Interactions between protein and ligands as seen in generated poses (orange) and re-docked poses (green) for (a) hydrogen bond acceptors and (b) hydrogen bond donors. **Generative models have significant difficulty making hydrogen bond interactions compared to the CrossDocked baseline** (shaded boxes). (c-d) Example of CrossDocked pose and generated pose, respectively.

approach we employ, and docking programs come with their inherent limitations, they contain useful proxies for comparison. We use SMINA [18] for all our experiments.

Experimental Setup In our study, we evaluate the quality of poses from seven recent methods: LiGAN [19], 3DSBDD [20], Pocket2Mol [7], TargetDiff [9], DiffSBDD [8], DecompDiff [11] and FLAG [10]. All models were trained on the CrossDocked2020 [21] dataset using the dataset splits computed in Peng et al. [7]. Further details on the CrossDocked test set, benchmarking and models implementation are given in Appendix B.

3 Results

3.1 Agreement with docking scoring functions

We first measure whether generated poses/binding modes from 3D SBDD models correspond to low energy states by computing the Root Mean Squared Deviation (RMSD) between the generated pose and the SMINA-minimized pose [18] in Figure 2. A larger RMSD would suggest that less information about the binding mode is preserved on minimization and there is less agreement with the scoring function. In short, we observe that none of the generative methods is able to generate poses in as low energy a state as the training data.

We first consider CrossDocked as a baseline, which has a mean minimization RMSD of 0.59 Å. 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 Å respectively. FLAG is the most egregious example with on average 3.64 Å RMSD during minimization and a maximum value of 10.72 Å, an extreme value for local minimisation.

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 [22], 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. Further discussion is provided in Appendix C.1.

3.2 Protein-ligand interaction analysis

We investigated the capacity of 3D SBDD generative models to create molecules with hydrogen bonding networking similar to those seen in the training datasets (Figure 3). Our findings reveal that none of the tested methods meets or surpass the baseline. In the baseline dataset CrossDocked, the

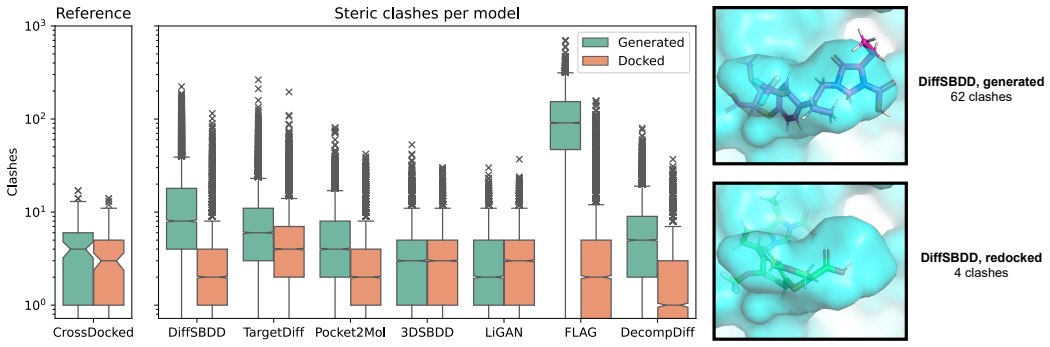

Figure 4: **Left:** number of steric clashes for the CrossDocked reference dataset as well as for the molecules generated by each model, both before and after re-docking. **Right:** examples of a generated pose (magenta) and the same pose after re-docking (green). **Diffusion and auto-regressive models exhibit more steric clashes compared to the baseline**

most frequent number of hydrogen bonds (HBs) was 1 for both acceptors and donors. The average numbers were 2.23 for acceptors and 1.66 for donors. In stark contrast, we found that for nearly all generated molecular structures from all models (with the exception of LiGAN's HB acceptors), the most common number of HB acceptors and donors was zero. We note that the creation of a HB requires a very specific geometry [23, 24], indicating these methods struggle to learn meaningful biophysics. Further details are elaborated in Appendix C.2.

3.3 Clash scores

Figure 4 presents the results of the steric clash analysis. In summary, the latest methods, particularly those employing diffusion models and auto-regressive fragment placement, exhibit poor performance in terms of steric clashes compared to the baseline, with a significant number of outliers. Although re-docking mitigates clashes to some extent, it does not always resolve the most severe cases, suggesting significant issues with the methods in terms of their ability to reason able protein-ligand complementarity. Notably, 3DSBDD and LiGAN show low clash scores, with the former positioning atoms within a predefined voxel grid [20] and the latter applying a clash loss [19]. DecompDiff also applies a steric clash loss (but does not directly measure clashes in the corresponding publication) [11] and performs best out of all the diffusion-based approaches. Further discussion is provided in Appendix C.3.

Interestingly, DiffSBDD and TargetDiff, both diffusion-based approaches [9, 8], 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 [20] and the latter applying a clash loss [19]. DecompDiff also applies a steric clash loss (but does not directly measure clashes in the corresponding publication) [11] 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 [10], regardless of whether it fits sterically.

3.4 Strain energy

To conclude our study, we provide an analysis of the strain energy [17] of the generated poses. Force field relaxation of the generated pose before docking is a common post-processing step of many generative SBDD pipelines, masking potential issues with the precise geometries of the generated molecules. Futhermore, high strain energy would be indicative that molecules are unlikely to bind.

In Figure 5a, we present the cumulative density function (CDF) of strain energy for all molecules generated by various models, using the CrossDocked dataset as a comparative baseline. Note that the x-axis is logarithmic. The data shows that most generative methods fall short of the CrossDocked's median strain energy of 3.96 kcal/mol, with FLAG and Pocket2Mol being the significant exceptions.

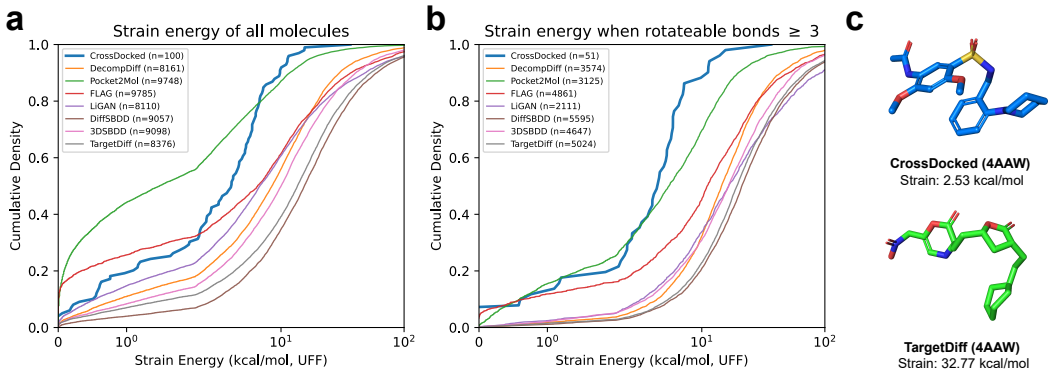

Figure 5: (a) CDF of strain energies for all molecules. (b) CDF of strain energy for all molecules that 3 or more rotatable bonds. **All of the generated molecules with 3 or more rotable bonds are more strained than the baseline**. (c) Example of TargetDiff generated molecule with strain substantially higher than in the CrossDocked test set (target 4AAW).

Upon examining the impact of rotatable bonds on strain energy, two key findings emerged: Firstly, molecules from generative models with a high count of rotatable bonds exhibit notably higher strain (refer to Appendix Figure 6). Secondly, the more successful methods in Figure 5a tend to produce molecules with fewer or no rotatable bonds, as detailed in Appendix Figure 7.

When focusing solely on molecules with three or more rotatable bonds, none of the methods surpassed the baseline performance (see Figure 5b). This discrepancy becomes more pronounced when assessing molecules with higher conformational complexity. Further details and extended results on this topic are available in Appendix C.4.

4 Recommendations for future work

Exploring reduced-noise sampling strategies Interestingly, both diffusion-based works (DiffSBDD and TargetDiff) performed similarly in terms of strain energy. We hypothesize this may be due to the injection of random noise into the coordinate features at all but the last step of stochastic sampling [25], making it challenging to construct precise bond angles. We recommend considering more sophisticated noising strategies that have been successful in protein design [26, 27].

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 [28] or the steric clash loss in LiGAN [19] and DecompDiff [11] (note that later method does note explicitly measure clashes).

Consider representing hydrogens Virtually all work in ML for structural biology chooses to not explicitly represent hydrogen atoms [28, 7, 20, 27, 8, 9], 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 [29] (even when redocked).

5 Conclusion

In conclusion, this work presents a comprehensive exploration of the limitations of structure-based drug design (SBDD) with deep generative models. In particular, we find that these methods have previously little understood failure modes about the quality of the 3D *poses* that these models produce jointly with the molecular topology. 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 and make a number of methodological recommendations. We provide PoseCheck as a solid evaluation suite and we hope that it stimulates further development towards more efficient drug discovery processes.

Acknowledgements

The authors would like to thank Jon Paul Janet, Alessandro Tibo, Juan Carlos Mobarec and Arne Scheuing for their valuable discussions related to this work. CH is supported by the Cambridge Centre for AI in Medicine Fellowship, which is in turn funded by AstraZeneca and GSK. ARJ acknowledges support from the Biotechnology and Biological Sciences Research Council (BBSRC) DTP studentship (BB/M011194/1).

References

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

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

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

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

[34] 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.

[35] 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.

[36] 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.

[37] J Michael Word, Simon C Lovell, Jane S Richardson, and David C Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, 285(4):1735–1747, 1999.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

A Background and Related Work

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) [5], Generative Adversarial Networks (GANs) [30], Autoregressive (AR) models and recently Denoising Diffusion Probabilistic Models (DDPMs) [4]. LiGAN [19] uses a 3D convolutional neural network combined with a VAE model and GAN-style training. 3DSBDD [20] 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 [7] extended this work with a more efficient sampling algorithm and better encoder. DiffSBDD [9], DiffBP [31] and TargetDiff [9] are all conditional DDPMs conditioned on the 3D target structure. DecompDiff [11] is another diffusion model that decomposes the ligand into fragments for which it considers separate priors for the diffusion process. FLAG [10] 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 [32] utilises an autoregressive flow model to formulate the ligand design as a sequential generation task.

Related work Guan et al. [9] 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. [33] 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 [11] and DiffBP [31] take steric clashes into account via their loss functions, but do not include steric clashes as a metric in their evaluation. TargetDiff [9] 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 [34] 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.

B Extended Implementation

B.1 CrossDocked Test Set

The CrossDocked dataset is a standard dataset used in the field of generative modelling for structure-based drug design [21]; 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 [35], 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 [18]. This cross-docking process results in the basis dataset CrossDocked 2020 [21], 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 [7]. 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 [36] 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. The 100 test samples consist of 28 hydrolases, 22 oxidoreductases, 11 kinases, 16 other transferases, 6 transcription factors, 4 lyases, 2 ligases, 1 GCPR, 2 membrane proteins, 2 isomerases, 1 viral protein, 1 transport protein, 1 signalling protein and 3 in other categories.

B.2 Benchmarking Methods Implementation

Protein-ligand interaction analysis Proteins are first protonated using Reduce [37] and interactions are then calculated using the ProLIF library [13].

Steric clashes 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. Proteins are protonated using Reduce [37] and molecules are protonated using RDKit[2].

Strain energy 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, $E_{\text{strain}} = E_{\text{generated}} - E_{\text{minimum}}$ as the difference between the estimated energy minimum and the energy of the generated pose (without pocket). Note, evaluating the generated poses with the force-field directly will cause the energy terms to explode, due to the slight imperfections in bond distances and angles in the generated molecules. Hence, we first perform at most 200 steps of relaxation using a force-feild with a maximum allowed displacement in atom positions of 0.1 Å. This fixes minor issues with the bond angles and distances, preventing the energy terms from exploding, whilst staying faithful to the global binding mode of the generated conformer. An estimate of the global energy minimum is calculated by initializing 50 conformers using ETKDGv3 and then minimizing using up to 200 steps of force-field minimization (taking inspiration from [34]). We then calculate the energy of all these poses and take the minimum as our final value. Both conformer minimization and energy evaluation are computed using the Universal Force Field (UFF) [38] using RDKit.

Docking We perform all pose scoring, minimization and redocking using SMINA [18]. Next, we can 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.

B.3 Experimental Setup

In our study, we evaluate the quality of poses from seven recent methods: LiGAN [19], 3DSBDD [20], Pocket2Mol [7], TargetDiff [9], DiffSBDD [8], DecompDiff [11] and FLAG [10]. All models were trained on the CrossDocked2020 [21] dataset using the dataset splits computed in Peng et al. [7], 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 B.1.

During inference, the model is given a reduced PDB file containing only the atoms for a single pocket within the test set, so there is no element of blind docking during generation or subsequent redocking[3].Docking protocols were done using the SMINA settings decribed in the original CrossDocked paper [21].

B.4 Procedure of Model Reproduction

For generated poses, we sourced molecules from Schneuing et al. [8] for DiffSBDD, and Guan et al. [9] for CrossDocked, TargetDiff, Pocket2Mol, 3DSBDD and LiGAN (where they provide generated poses but we additionally perform our own redocking).

For FLAG [10], no weights were provided so we retrained the model as described in Zhang et al. [10] 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.

[2] `www.rdkit.org`

[3] Note illustrative figures may show full proteins.

For DecompDiff [11], 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 Extended Results

C.1 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 [18], with a lower RMSD value denoting a higher degree of agreement.[4]

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. 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 autoregressive models, 3DSBDD and Pocket2Mol, which obtain mean RMSDs of 0.99 and 1.02 Årespectively. 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.

Table 1: Vina score values of generated poses, energy minimized poses and redocked poses. We additionally provide the change in Vina score during minimization and redocking respectively.

Method	**Vina Generated (↓)** (kcal/mol)	**Vina Minimized (↓)** (kcal/mol)	**Vina Redocked (↓)** (kcal/mol)	**ΔAffinity Minimization (↑)** (kcal/mol)	**ΔAffinity Redocking (↑)** (kcal/mol)
CrossDocked	-5.50 ± 2.86	-6.24 ± 2.52	-6.86 ± 2.37	-0.74 ± 1.24	-1.37 ± 1.43
TargetDiff	-5.36 ± 3.79	-6.72 ± 2.83	-7.35 ± 2.51	-1.35 ± 1.99	-1.99 ± 2.59
3DSBDD	-5.04 ± 2.58	-5.85 ± 2.42	-6.29 ± 2.22	-0.80 ± 0.76	-1.25 ± 1.0
Pocket2Mol	-4.55 ± 3.18	-6.38 ± 2.92	-6.96 ± 2.72	-1.83 ± 1.66	-2.40 ± 2.01
DecompDiff	-4.25 ± 3.16	-5.91 ± 2.14	-6.56 ± 2.03	-1.66 ± 2.39	-2.31 ± 2.69
LiGAN	-6.03 ± 2.83	-6.78 ± 2.71	-7.36 ± 2.56	-0.75 ± 0.78	-1.34 ± 1.09
DiffSBDD	-1.94 ± 10.31	-5.85 ± 3.19	-7.00 ± 2.01	-3.91 ± 8.62	-5.07 ± 9.92
FLAG	94.20 ± 89.46	4.89 ± 19.36	-5.69 ± 4.19	-89.31 ± 78.45	-99.89 ± 88.94

We also provide the raw affinities from our docking experiments in Table 1, both when evaluating the generated pose using the SMINA/Vina score function directly (Vina Generated), after local energy minimization (Vina Minimize) and redocking the molecule entirely (Vina Redock). Additionally, we provide the change in Vina scores during minimization and redocking. We first draw our attention to the scores for the redocked poses: these metrics are commonly reported and often used as a justification for state-of-the-art performance. On the surface, the results look promising, with most methods matching or exceeding the performance of the baseline dataset (although with no statistical significance). However, a worrying picture emerges when we measure the generated poses directly, with none of the models (except LiGAN) outperforming the baseline dataset. The mean scores for CrossDocked are -5.50 kcal/mol, whereas the generative models (except LiGAN and FLAG) have mean scores between -1.94 and -5.36 kcal/mol. FLAG again performs poorly with the Vina scores for generated poses exploding to +94.20 kcal/mol, suggesting the generated poses are highly implausible.

We next consider the role energy minimization and redocking have on these final scores by considering the change in affinity during the two processes respectively. The result of this analysis highlights that minimization/redocking is critical to getting acceptable scores out of these methods, calling into question the reliability of the generated poses. CrossDocked has a Δaffinity minimization score of -0.74 kcal/mol, whereas the generative models (excluding LiGAN and FLAG) have scores between -0.75 and -3.91 kcal/mol. FLAG has a score of -89.31 kcal/mol (unsurprising given the generated poses). We see a similar picture for the impact of redocking, where the majority of methods see

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

substantially greater increases in their scores during the procedure. In conclusion, we show that only reporting the mean Vina score of redocked poses hides critical failure modes found in many models. [5]

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 [22], 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. [21] provided up to 20 poses for every ligand, resulting in 22.5 million complexes, and the processing done in Peng et al. [7] is not clear on which poses they chose, meaning these models may not have been trained on the lowest affinity poses.

C.2 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 [23]. They are key to many protein-ligand interactions [24] and require very specific geometries to be formed [29]. 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 [39]. However, they can be significant when many atoms are involved, as is typical in protein-ligand binding [40]. **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 [41], thereby facilitating the association between protein and ligand molecules [42].

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 8), 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.

C.3 Clash scores

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.

[5] Given the acceptable redocked scores of FLAG we do not believe we have made an error in training.

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 [9, 8], 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 [20] and the latter applying a clash loss [19]. DecompDiff also applies a steric clash loss (but does not directly measure clashes in the corresponding publication) [11] 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 [10], 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 [28]. 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 [43]. 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.

C.4 Strain energy

The additional analysis of the impact of rotatable bonds on strain and the frequency of rotatable bonds are shown in Figure 6 and Figure 7 respectively.

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 8.

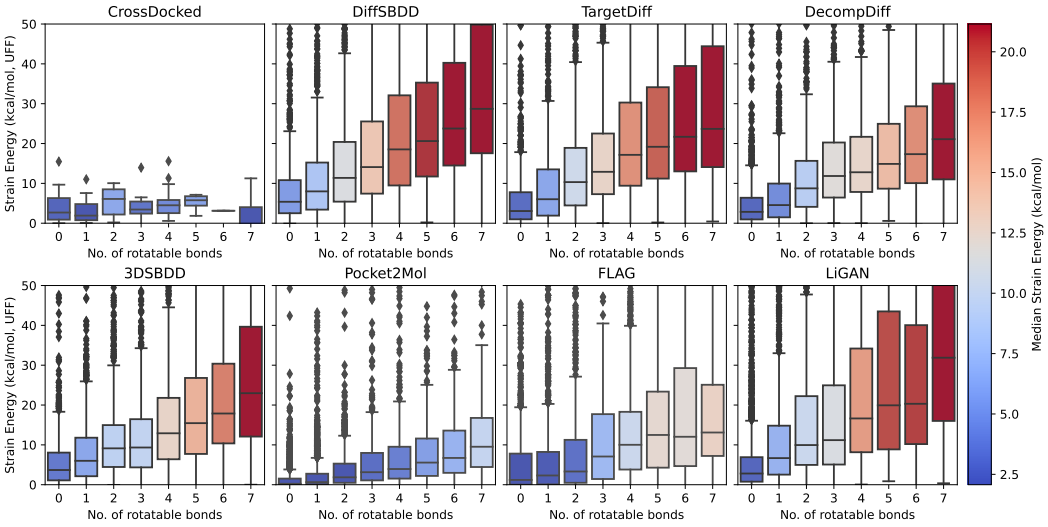

Figure 6: Boxplots of strain energies of generated molecules per number of rotatable bonds for all methods. Box color shows median strain value.

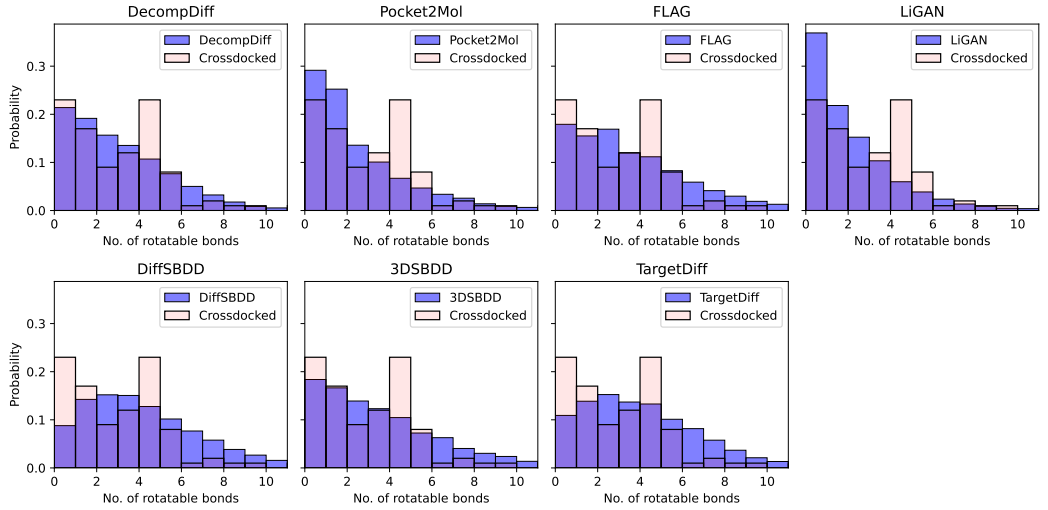

Figure 7: Histograms showing distributions of rotatable bonds for all molecules generated by a particular method. In each plot, the underlying distribution from CrossDocked is also shown.

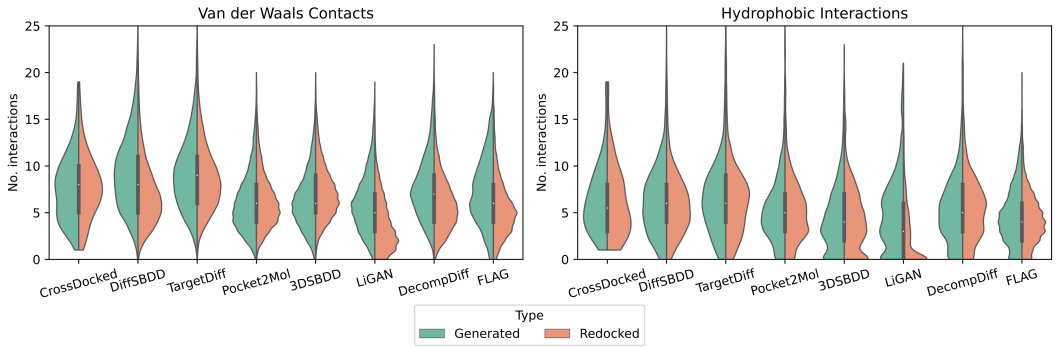

Figure 8: Extended analysis of the interaction profiles of the generated molecules for the different methods. While the focus in the main text was on hydrogen bonds, the results in this figure include Van der Waals Contacts and hydrophobic interactions, reported for both the generated as well as the redocked pose.