Structure-Conditioned Generative Models for De Novo Ligand Generation: A Pharmacophore Assessment

Shannon T. Smith,[1] **Leo Gendelev,**[2] **Kangway V. Chuang,**[3†] **Seth F. Harris**[1†]
[1]Department of Structural Biology,
[2]Department of Discovery Chemistry
[3]Department of Machine Learning for Drug Discovery
Genentech Research and Early Development
[†]{chuang.kangway,harris.seth}@gene.com

Abstract

Deep generative models show promise for de novo molecular design, especially pocket-conditioned conditional generation methods that output small-molecule ligands in their predicted binding pose with high shape complementarity. However, recent work demonstrates these models still fail to generate chemically valid and synthetically accessible ligands. This paper provides further insight into these methods and their generated molecules through analysis of pharmacophore features commonly used in structure-based and ligand-based drug discovery. We specifically assess the generated distribution of hydrogen bond donors, acceptors, and aromatic rings from deep generative methods on three well-studied protein targets: adenosine A2a receptor, cyclin-dependent kinase 2, and the main protease of SARS-CoV-2. Our results find autoregressive approaches better recapitulate the expected spatial distribution of pharmacophore features compared to diffusion-based models. The analysis presented here highlights current limitations in deep generative models for 3D design, while suggesting new directions to realistically aid structure-based design.

1 Introduction

The ability to accurately predict how drug candidates bind to protein targets is critical for successful structure-based drug design. Traditionally, this involves exhaustive conformational sampling and molecular docking, which is computationally demanding and restricted to existing molecular libraries [1, 2]. Recently, deep generative models have unlocked exciting possibilities for de novo molecular design. By learning to produce novel ligands conditioned directly on binding sites, these models bypass limitations of predefined chemical spaces and separate docking steps. Architectures utilizing three-dimensional diffusion or autoregressive generation have shown initial promise in producing diverse ligands with favorable shape complementarity. Moreover, jointly generating candidate structures and bound poses is highly appealing to streamline drug discovery [3].

Recent studies have underscored limitations in deep generative models for structure-based drug design (SBDD) through the introduction of new pose evaluation methodologies [4, 5]. In particular, generated molecules were found to violate validity and physical constraints and provide lower-quality poses than traditional docking software [6]. Together, these analyses provide frameworks to pinpoint deficiencies and drive future enhancements through rigorous pose evaluation. Although recent benchmarks have evaluated overall ligand quality [7] and physical validity, an unanswered question is whether deep generative SBDD models are effectively learning meaningful representations of protein-ligand interactions essential for binding. In contrast to prior approaches that strictly assess ligands and their interactions, we instead take an abstracted approach by extracting pharmacophore features, e.g. hydrogen bond acceptor (HBA), hydrogen bond donor (HBD), etc., from generated

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

poses. Relaxing the metric through "fuzzy-matching" of pharmacophores provides a useful, but less-stringent, evaluation of model quality, by using these features to understand spatial patterns. By evaluating pose pharmacophore fidelity, our work provides an interpretative lens to guide refinements in deep generative SBDD. We summarize our contributions as follows:

- We qualitatively analyze several state-of-the-art deep generative models, 3DGenSBDD, Pocket2Mol, and DiffSBDD across multiple protein targets, adenosine A2a receptor, cyclin-dependent kinase 2, and the main protease of SARS-CoV-2. We demonstrate that the spatial distributions of key pharmacophore features tend to be better localized with respect to idealized pharmacophore features in autoregressive approaches compared to diffusion-based approaches.
- We develop a simple coarse clustering analysis of ligand pharmacophore features to assess whether models learn to consistently place molecular features, and find greater dispersion with diffusion models across different protein families.
- Leveraging these insights, we suggest several key directions for further exploration in this area to improve model training.

2 Background and Related Work

Small Molecule Generative Models. Earlier approaches focused on generating 1D SMILES strings [8] or 2D molecular graphs. For example, GraphAF [9] leverages normalizing flows on molecular graph representations. Jin et al. [10] use a junction tree variational autoencoder that operates on molecular graphs. While powerful, these methods do not model 3D geometry explicitly. More recent techniques [11] aim to generate molecular structures directly in 3D space . Pocket2Mol [12] uses an autoregressive model that sequentially adds atoms and bonds based on a 3D protein pocket context. Li et al. [13] frame the generation problem as a reinforcement learning task and use Monte Carlo tree search to generate pocket-specific ligands. However, sequential atom-by-atom generation can be inefficient and may not capture global 3D relationships.

Structure-Based Diffusion Models. Building on above advances, diffusion models have been adapted for structure-based molecular generation tasks. DiffSBDD [14] employs an SE(3)-equivariant diffusion model to generate novel ligands conditioned on protein binding pockets. They demonstrate both context encoding and ligand inpainting strategies. While showing strong promise, current models still have limitations; they assume static protein structures and cannot model induced fit or significant flexibility. Exploring conditioning strategies and architectures to address this remains an open challenge. Our work seeks to evaluate these models and whether they learn realistic binding features.

Benchmarking 3D Molecular Generative Models Harris et al. [5] presented POSECHECK to assess binding poses from several state-of-the-art SBDD methods (including LiGAN, 3D-GenSBDD, Pocket2Mol, TargetDiff [15], and DiffSBDD). Despite leveraging 3D structural information, the generated molecules frequently violated physical constraints and formed fewer key receptor interactions than expected, calling into question presumed benefits of explicit 3D conditioning. In a complementary study, Buttenschoen et al. [4] developed PoseBusters, a test suite benchmarking both deep learning and classical docking approaches. Testing five deep generative models (DeepDock [16], DiffDock [17], EquiBind [18], TankBind, UniMol) and two traditional docking tools (AutoDock Vina [19], Gold [20]), they demonstrated key shortcomings in deep learning approaches, which failed to surpass classical programs in producing chemically valid and favorable binding poses, especially for novel protein targets.

3 Methods

3.1 Evaluated Models

We evaluated several diverse generative models that condition based on a protein pocket that simultaneously generate a ligand and its pose:

3DGenSBDD Luo et al. [21] recently develop a 3D generative model to estimate the probability density of atom occurrences in a protein's binding site. The model uses a graph neural network

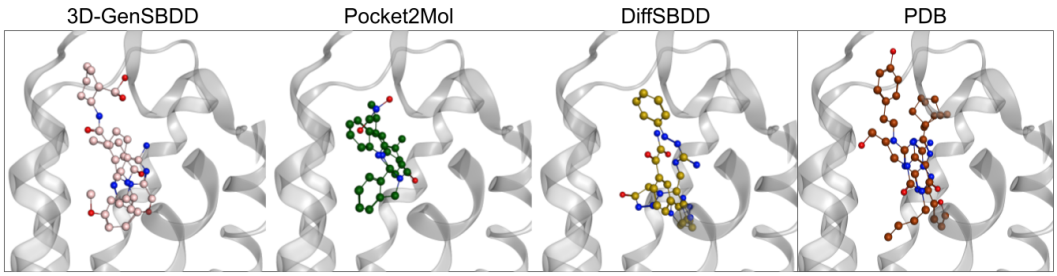

Figure 1: Two example molecules from each method assessed.

encoder composed of rotationally invariant message passing layers to represent the 3D context. An autoregressive approach is used to generate valid and diverse small molecules from the learned distribution by sequentially sampling atoms based on predictions from a spatial classifier module.

Pocket2Mol Peng et al. [12] introduce an E(3)-equivariant network for molecular sampling conditioned on 3D protein pockets. It employs a graph neural network with vector-based neurons and geometric vector perceptrons to model the chemical and geometric features of the pockets. Atoms, types and bonds are efficiently generated by directly predicting relative atom positions from tractable distributions without relying on MCMC sampling.

DiffSBDD Schneuing et al. [14] disclose an SE(3)-equivariant 3D conditional diffusion model for structure-based drug design. It leverages equivariant graph neural networks to generate novel ligands conditioned on protein binding pockets. The model can be trained in a protein-conditional or joint manner, and applied to downstream tasks like property optimization and scaffold hopping.

3.2 Protein Targets and Sampling

We study these models in the context of three structurally well-studied protein targets: adenosine A2a receptor 2 (A2a, PDB: 5NM4 [22]), cyclin-dependent kinase 2 (CDK2, PDB: 6Q4H [23]), and SARS-CoV-2 main protease (Mpro, PDB: 5R82 [24]) were chosen due to a large number of publicly-deposited structures with bound, non-covalent ligands in a pocket of interest; the highest resolution structure was chosen for generation. For each method on each target, we generated 10,000 molecules. Each resulting molecule was run through RDKit to obtain Cartesian coordinates for hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), aromatic ring centroid (ARO), cationic (CAT), anionic (ANI), and hydrophobic (HYP) atoms. We also generated coordinates for Van der Waals (VDW) and ligand centroids (CTR) based on all heavy atoms. Protein pharmacophore features were extracted and visualized using the Molecular Operating Environment (MOE).

3.3 Clustering Assessment

Clustering assessment was performed using Density-based spatial clustering of applications with noise (DBSCAN) [25] using the scikit learn package. We looked at the number of clusters, percentage of noise points where no cluster was assigned, and the Calinski-Harabasz index as a measure of cluster definition where higher values indicate better-defined clusters. To ensure a consistent number of samples for DBSCAN-based clustering, we extracted 3,000 coordinates for each pharmacophore feature group.

4 Results and Discussion

Whereas prior studies assessing generative models [4, 5] focus on ligand quality or assess their steric clashes, we focus instead on the question of *feature localization*, *i.e.* do generative models tend to place chemical features for a generated ligand that complement the protein pocket? To address this question, we visualize features from a large pool of generated ligands, assess their radial placement with respect to key interacting residues, and assess their clustering within the pocket. As illustrated in Fig. 2 and Table 1, autoregressive vs. diffusion-based approaches generate ligands in dramatically different ways. The far right panel illustrates the observed PDB reference ligands that

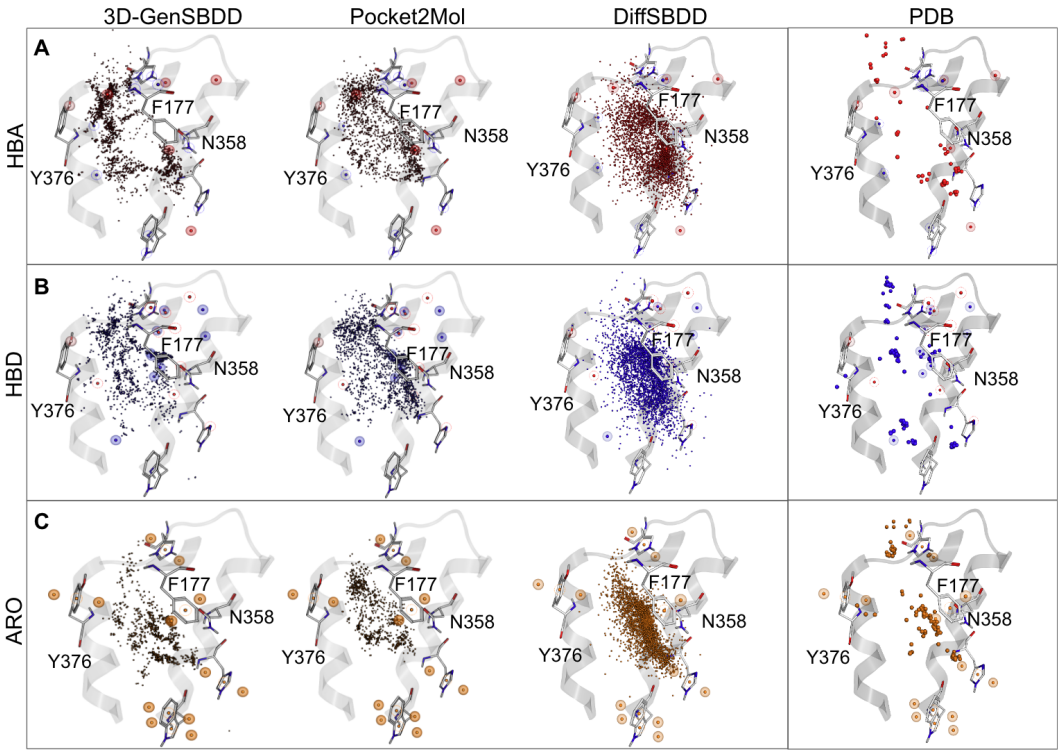

Figure 2: Overlay of generated ligand pharmacophore features (small dots) with MOE pharmacophore feature projections (larger spheres) from receptor atoms in A2a pocket. A) ligand HBAs overlaid on projected HBA coordinates (filled red spheres) based on receptor HBDs (empty blue spheres), B) ligand HBDs overlaid on projected HBD coordinates (filled blue spheres) based on receptor HBAs (empty red spheres), C) ligand aromatic ring centroids overlaid on projected pi-normal projections (filled orange spheres) based on receptor aromatic rings. Each image contains 3k coordinates from 10k generated molecules.

satisfy key hydrogen bonding and aromatic interactions to key residues N358 and F177. We find that the autoregressive models tested place molecular features more systematically around the pocket, whereas diffusion-based approaches largely focus on pocket occupancy.

We also reported radial distributions of generated ligand pharmacophore features (Figure 3) from three receptor atoms critical for PDB ligand interactions: N358 sidechain nitrogen to ligand HBAs, N358 sidechain carbonyl oxygen to ligand HBDs, and the F177 aromatic centroid to ligand aromatic

Table 1: Clustering results for generated atoms against the A2a receptor.

Feature	Metric	3D-GenSBDD	Pocket2Mol	DiffSBDD
HBA	clusters	11	19	4
	noise points	234	342	926
	sil. coeff.	0.021	-0.112	-0.216
	C–H index	497	234	25
HBD	clusters	11	20	2
	noise points	259	373	754
	sil. coeff.	-0.014	-0.130	0.056
	C–H index	220	212	35
Aromatic	clusters	9	2	1
	noise points	87	77	305
	sil. coeff.	0.286	-0.032	0.330
	C–H index	474	15	13

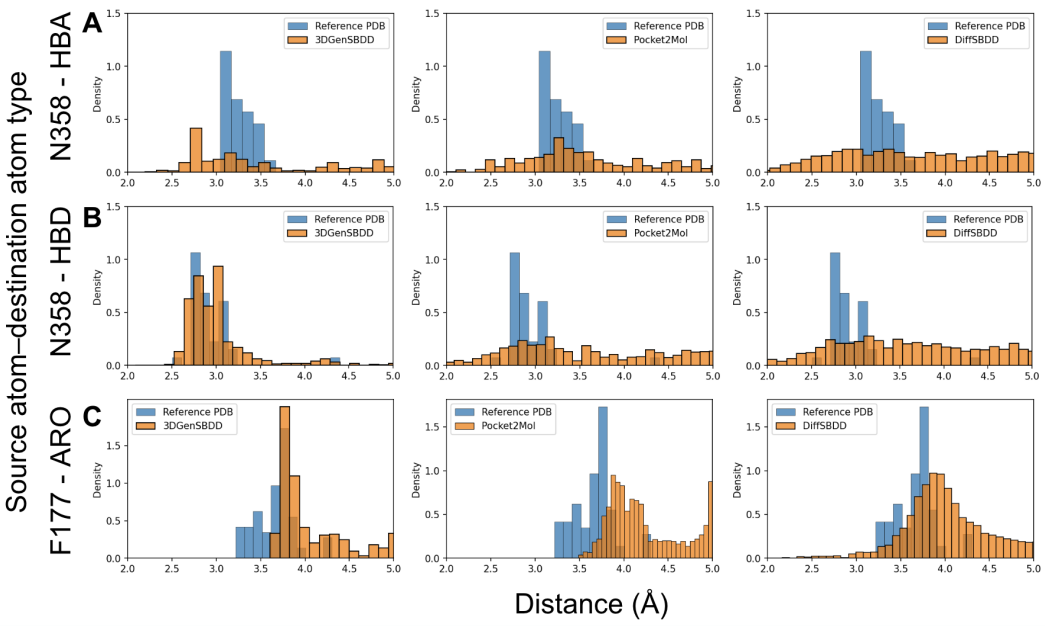

Figure 3: Histograms of distances between generated atomic coordinates for hydrogen bond acceptors (HBAs), donors (HBDs), and aromatic rings (ARO) from key pharmacophore residues (N358, N358, F177) from A2a binding pocket. Generated distances are represented in orange bars. Observed radial distances for reference PDB ligands are shown in the blue bars. Molecular interactions typically satisfy narrow ranges, which are not well captured by most models.

centroids. PDB ligands expectedly show tight HBA and HBD distances between 2.7-3.5Å and aromatics between 3.5-4.0Å. HBAs somewhat favor this distance for the autoregressive models, but a more uniform distribution is observed for DiffSBDD. HBDs for 3D-GenSBDD overlay tightly with the HBD distance of PDB ligands, whereas we don't observe this same effect for Pocket2Mol and DiffSBDD. Moreover, these methods do not generate aromatic rings <3.5Å, and these each show peaks near the 3.8Å. Furthermore, because ligand binding requires satisfying key protein-ligand interactions, we further explored the hypothesis that generated ligands points should generate natural pharmaocphore feature clusters. To test this hypothesis, we performed clustering using DBSCAN for each method to identify potential clusters, and observed distinct modes for 3D-GenSBDD and Pocket2Mol, but noted high dispersion for DiffSBDD. Although these analysis focus on more qualitative insights, we believe they provide direct insight into the extent that models can satisfy critical physical interactions, rather than simply filling the pocket.

Given these observations, we propose that future development in this active research area leverage pharmacophore-based evaluations for model assessment and improvement. Satisfying key protein-ligand interactions remain key in the optimization of small molecules; integrating this domain knowledge into model design may further improve model performance. In particular, we propose: 1) Incorporate chemical and geometric intuition to guide generation, rather than purely data-driven approaches. Domain knowledge may help satisfy critical interactions. 2. Develop multi-objective frameworks that balance competing factors like novelty, diversity, and interaction fidelity during training and sampling to provide a more holistic approach. 3. Explore constraints and losses to enforce correct pharmacophore matching and geometry to improve satisfaction of key interactions. We anticipate that exploration of these approaches, in concert with complementary approaches proposed by Harris et al. [5] and Buttenschoen et al. [4] will yield significant improvements. All together, we believe that further research in this area will continue to improve on these early approaches towards more realistic and useful molecule and pose generation.

References

[1] Brian K Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–865, December 2004.

[2] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A Tolmachev, Brian K Shoichet, Bryan L Roth, and John J Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019.

[3] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist's guide to molecular interactions. *J. Med. Chem.*, 53(14):5061–5084, jul 2010.

[4] 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, 2023.

[5] Charles Harris, Kieran Didi, Arian R. Jamasb, Chaitanya K. Joshi, Simon V. Mathis, Pietro Lio, and Tom Blundell. Benchmarking generated poses: How rational is structure-based drug design with generative models?, 2023.

[6] Tobiasz Ciepliński, Tomasz Danel, Sabina Podlewska, and Stanisław Jastrzbski. Generative models should at least be able to design molecules that dock well: A new benchmark. *Journal of Chemical Information and Modeling*, 63(11):3238–3247, may 2023. doi: 10.1021/acs.jcim.2c01355. URL https://doi.org/10.1021%2Facs.jcim.2c01355.

[7] Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Günter Klambauer. Fréchet ChemNet distance: A metric for generative models for molecules in drug discovery. *Journal of Chemical Information and Modeling*, 58(9):1736–1741, aug 2018. doi: 10.1021/acs.jcim.8b00234. URL https://doi.org/10.1021%2Facs.jcim.8b00234.

[8] 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 Cent Sci*, 4(2):268–276, February 2018.

[9] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang. Graphaf: a flow-based autoregressive model for molecular graph generation, 2020.

[10] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation, 2019.

[11] Clemens Isert, Kenneth Atz, and Gisbert Schneider. Structure-based drug design with geometric deep learning. *Current Opinion in Structural Biology*, 79:102548, 2023. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2023.102548. URL https://www.sciencedirect.com/science/article/pii/S0959440X23000222.

[12] Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2Mol: Efficient molecular sampling based on 3D protein pockets. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17644–17655. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/peng22b.html.

[13] Yibo Li, Jianfeng Pei, and Luhua Lai. Structure-based de novo drug design using 3d deep generative models. *Chem. Sci.*, 12:13664–13675, 2021. doi: 10.1039/D1SC04444C. URL http://dx.doi.org/10.1039/D1SC04444C.

[14] Arne Schneuing, Yuanqi Du, Charles Harris, Arian Rokkum Jamasb, Ilia Igashov, weitao Du, Tom Leon Blundell, Pietro Lio, Carla P Gomes, Max Welling, Michael M. Bronstein, and Bruno Correia. Structure-based drug design with equivariant diffusion models, 2023. URL https://openreview.net/forum?id=uKmuzIuVl8z.

[15] 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, 2023.

[16] Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, dec 2021. doi: 10.1038/s42256-021-00409-9. URL https://doi.org/10.1038%2Fs42256-021-00409-9.

[17] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kKF8_K-mBbS.

[18] Hannes Stärk, Octavian-Eugen Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction, 2022.

[19] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, jul 2021. doi: 10.1021/acs.jcim.1c00203. URL https://doi.org/10.1021%2Facs.jcim.1c00203.

[20] Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, aug 2003. doi: 10.1002/prot.10465. URL https://doi.org/10.1002%2Fprot.10465.

[21] Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 6229–6239. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/314450613369e0ee72d0da7f6fee773c-Paper.pdf.

[22] Tobias Weinert, Natacha Olieric, Robert Cheng, Steffen Brunle, Daniel James, Dmitry Ozerov, Dardan Gashi, Laura Vera, May Marsh, Kathrin Jaeger, Florian Dworkowski, Ezequiel Panepucci, Shibom Basu, Petr Skopintsev, Andrew S Dore, Tian Geng, Robert M Cooke, Mengning Liang, Andrea E Prota, Valerie Panneels, Przemyslaw Nogly, Ulrich Ermler, Gebhard Schertler, Michael Hennig, Michel O Steinmetz, Meitian Wang, and Jorg Standfuss. Serial millisecond crystallography for routine room-temperature-structure determination at synchrotrons. *Nat. Commun.*, 8(1):542, sep 2017.

[23] Daniel J. Wood, J. Daniel Lopez-Fernandez, Leanne E. Knight, Islam Al-Khawaldeh, Conghao Gai, Shengying Lin, Mathew P. Martin, Duncan C. Miller, Céline Cano, Jane A. Endicott, Ian R. Hardcastle, Martin E. M. Noble, and Michael J. Waring. Fraglites—minimal, halogenated fragments displaying pharmacophore doublets. an efficient approach to druggability assessment and hit generation. *Journal of Medicinal Chemistry*, 62(7):3741–3752, 2019. doi: 10.1021/acs.jmedchem.9b00304. URL https://doi.org/10.1021/acs.jmedchem.9b00304. PMID: 30860382.

[24] Alice Douangamath, Daren Fearon, Paul Gehrtz, Tobias Krojer, Petra Lukacik, C. David Owen, Efrat Resnick, Claire Strain-Damerell, Anthony Aimon, Péter Ábrányi-Balogh, José Brandão-Neto, Anna Carbery, Gemma Davison, Alexandre Dias, Thomas D. Downes, Louise Dunnett, Michael Fairhead, James D. Firth, S. Paul Jones, Aaron Keeley, György M. Keserü, Hanna F. Klein, Mathew P. Martin, Martin E. M. Noble, Peter O'Brien, Ailsa Powell, Rambabu N. Reddi, Rachael Skyner, Matthew Snee, Michael J. Waring, Conor Wild, Nir London, Frank von Delft, and Martin A. Walsh. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature Communications*, 11(1), oct 2020. doi: 10.1038/s41467-020-18709-w. URL https://doi.org/10.1038%2Fs41467-020-18709-w.

[25] M Ester, H P Kriegel, J Sander, and Xu Xiaowei. A density-based algorithm for discovering clusters in large spatial databases with noise. 12 1996. URL https://www.osti.gov/biblio/421283.

Appendix

A Additional Plots

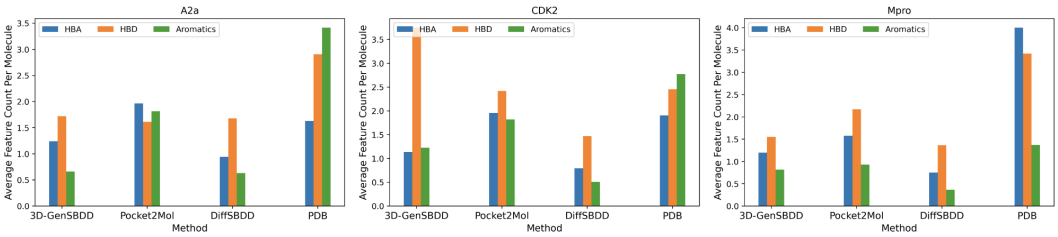

Figure 4: Average composition of pharmacophore features across the three targets tested compared to known PDB ligands.

In general, each method generates less HBA, HBD and ARO groups on average compared to ligands deposited in the PDB. As PDB ligands are often highly optimized, this is not surprising; however, it may emphasize the need for these methods to upweight these groups during generation.

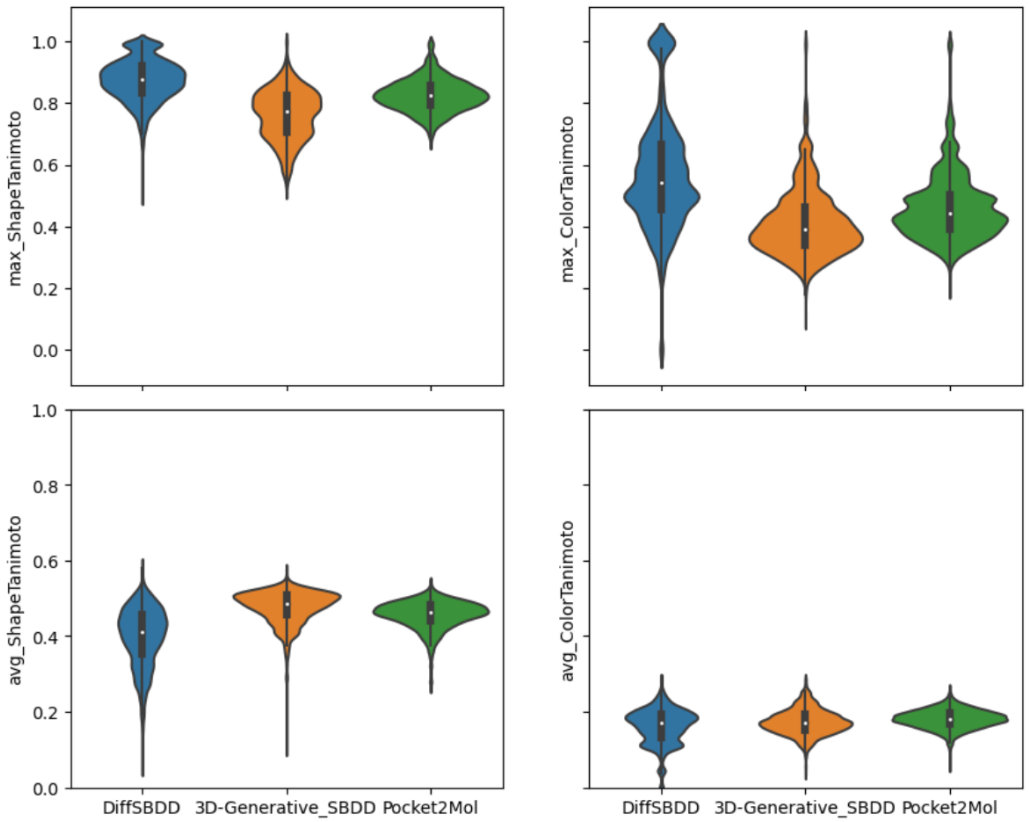

Figure 5: Distributions of Mpro generated molecule properties compared to PDB reference ligands. For each generated molecule, we overlaid this onto every PDB ligand without re-optimization and calculated the shape and color Tanimoto scores using OpenEyeOmega's FastROCS. The top represents the maximum Tanimoto similarity for each generated ligand to any PDB reference ligand; the bottom represents the average Tanimoto similarity for each generated ligand across all PDB reference ligands. Left is shape-based Tanimoto; right is color-based.

We used the FastROCS analysis to investigate how well generated molecules overlap to PDB ligands, both sterically (shape) and chemical (color) features. Generated molecules are in the same coordinate space as PDB ligands, and no optimization was performed.

For the max scores (top), the best overlap for each generated molecule is shown here. In all methods, the best steric overlap between 0.8-0.9 suggests that these methods successfully fill similar parts of the pocket to PDB ligands. These max scores drop significantly in the chemical features overlap, where maximum Tanimoto scores achieve 0.5 on average. Interestingly, there is a bimodal distribution in DiffSBDD; however, as we have seen that these features uniformly fill the pocket, we would assume that some of these fall close to PDB ligands. The average Tanimoto scores of each generated molecule (bottom) show that on average, these molecules do not align closely, either by shape or chemical features. One specific observation is that no method extended into the solvent-facing regions, a strategy common among medicinal chemists. It may be advantageous for these methods to recognize bulk water and build hydrophilic moieties accordingly.