Ensemble Guidance: Towards Generative 3D SBDD in Bioactive Chemical Spaces

Anonymous Authors[1]

Abstract

Many works use diffusion generative modelling for 3D Structure-based Drug Design. The data these models are trained on are predominantly sourced from the Protein Data Bank (PDB); these datasets capture a severely constrained and skewed subset of chemical space, heavily biasing generated molecules to be non-drug like whilst significantly narrowing the diversity of the chemical landscapes generative models observe during training. While there is some evidence these methods can generate complimentary molecules, this raises concerns about efficacy in novel hit discovery compared to virtual screening of large molecule libraries. Here, we introduce ensemble guidance, a technique for composing learned distributions from multiple diffusion models to guide SBDD models to generate molecules with more appropriate properties and higher diversity. For example, ensemble guidance reduces the frequency of highly polar phosphate groups from 0.32 per molecule to 0. Finally, we propose many areas of future work and hope that ensemble guidance can be fruitfully applied to a number of other (bio)molecular design tasks in data-limited regimes.

1. Introduction

Structure-based drug design (SBDD) is the task of designing a small molecule compounds that binds a protein receptor selectively (Blundell, 1996). This problem is difficult due the large design space, estimated to be 10^{60} molecules in size (Polishchuk et al., 2013). Traditionally, this design space has been searched by virtual screening of large libraries for diverse compounds/scaffolds to find starting points, called hits, which are further optimized to make suitable drugs (Keserű & Makara, 2006). Recently the machine learning community has proposed several methods attempting to perform SBDD using 3D generative modelling (Schneuing et al., 2022; Peng et al., 2022; Torge et al., 2023; Drotár et al., 2021). These methods are usually trained conditionally on paired data of protein-ligand complexes from either experiments or docking calculations, and then attempt to generate new hits for a given protein binding pocket. These methods are well-known to be far from perfect, particularly demonstrating severe limitations in the synthetic accessibility of designs (Gao & Coley, 2020) and limited physical plausibility of the generated poses (Harris et al., 2023a).

We suggest that these methods are, in part, significantly limited by the diversity and quantity of crystallographic training data. Plainly stated, the Protein Data Bank (PDB) contains protein-ligand complexes for molecules that structural biologists have studied for purposes often unrelated to drug discovery. Many of the ligands exhibit limited suitability in drug discovery campaigns, which in turn results in generated molecules unsuitable for drug discovery or development (e.g. many polar groups). Meanwhile, ultra-large chemical libraries offer an attractive alternative through the diversity of the compounds present (Fig 1). Here, we propose *ensemble guidance*, a method for composing multiple generative models to guide SBDD models to improve the drug-likeness of designs. We summarise our contributions as:

1. We show the ligands commonly used for training SBDD models from paired protein-ligand complex datasets are extremely biased, lack diversity, and over-represent motifs/functional groups unsuitable for drug discovery in comparison to larger screening libraries.

2. We introduce a method from composing the score functions learned by an ensemble of separate diffusion models trained on different kinds and scales of datasets, improving design diversity and drug-likeness, whist maintaining protein-ligand complementary.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the AI for Science Workshop at ICML 2024. Do not distribute.

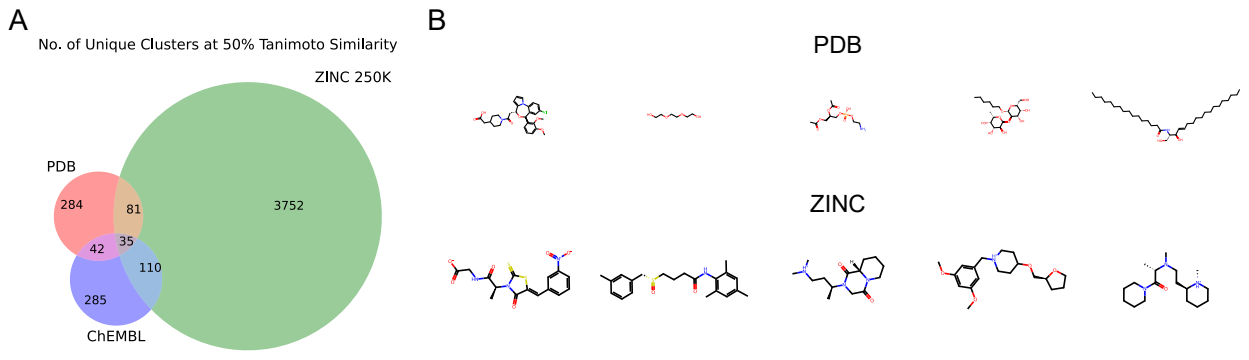

Figure 1. (A) Disparity between ligands in the PDB and those found in larger chemical spaces. Numbers in Venn diagrams represent the number of unique molecule clusters in that space. Molecular fingerprints were clustered to 50% Tanimoto similarity using Butina clustering. (B) Cluster centroid molecules from the top five largest clusters for the PDB and ZINC 250K ligands. Note the poor drug-likeness of molecules drawn from the PDB.

1.1. Background

Diffusion Score-based Models (DSMs) DSMs (Ho et al., 2020) are a class of latent variable model that learn a data distribution $p(x)$ by approximating the *score function*, that is the gradient of the log probability density $\nabla_x \log p(x)$. This approach allows DSMs to iteratively refine samples, guiding them towards high-probability regions of the data distribution, following a process that can be modeled via a Stochastic Differential Equation (Song & Ermon, 2019):

$$dx = \left[f(x,t) - g^2(t)\nabla_x \log p_t(x)\right] dt + g(t)dW.$$

where dW is a Wiener process. Through a process akin to reverse diffusion, these models gradually denoise data, starting from a random noise distribution and progressively converge to the data distribution. The training of DSMs involves optimizing a denoising score matching objective, which encourages the model's estimated score to align with the true score of the data at various noise levels.

3D SBDD with Generative Models 3D SBDD and Pocket2Mol (Peng et al., 2022) build molecules by autoregressively generating molecules atom-by-atom, while FLAG (Zhang et al., 2023) generates based on fragments. DiffSBDD (Schneuing et al., 2022) and TargetDiff (Guan et al., 2023) use a conditional diffusion models and can been seen as an conditional extension of the Equivariant Diffusion Model (EDM) (Hoogeboom et al., 2022). DiffLinker (Igashov et al., 2022) and DiffHopp (Torge et al., 2023) are specialised models for fragment linking and scaffold hopping respectively while Harris et al. (2023b) showed that a pretrained diffusion model can be adapted at sampling time for accomplish a number of tasks.

Classifier-free guidance Classifier-free guidance (Ho & Salimans, 2022) is a technique for conditioning diffusion models with certain labels without the need for an explicit classifier. By interpolating between the gradients of the log probabilities of the conditional and unconditional data distributions, the model can be guided towards generating samples that satisfy the desired conditions. The interpolation is controlled by a parameter γ, which can be tuned to adjust the strength of the conditioning. The equation below formalizes this concept:

$$\nabla \log p(x_t|y) = \gamma \underbrace{\nabla \log p(x_t|y)}_{\text{conditional score}} + (1-\gamma) \underbrace{\nabla \log p(x_t)}_{\text{unconditional score}}$$

This method allows biasing of the generation process towards certain attributes, effectively guiding the diffusion process.

2. Ensemble guidance

2.1. Datasets

PDB We use a subset of protein-ligands complexes in the PDB as processed by Brocidiacono et al. (2023)[1] As they are experimental, these complexes can be seen as the gold standard for training any structure-conditioned model. Receptors are split based on pocket similarity using ProBiS (Konc & Janežič, 2010); we use a subset of 100 proteins from the test set due to computational limitations.

[1]Note these were originally processed in the CrossDocked (Francoeur et al., 2020) method but Brocidiacono et al. (2023) only use experimentally-determined complexes and contains no docked poses.

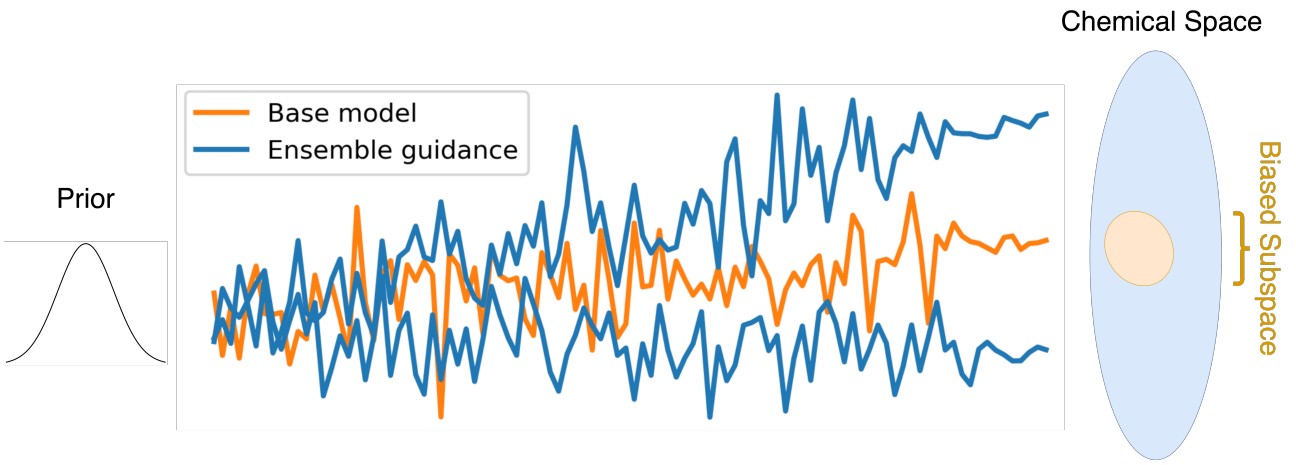

Figure 2. Ensemble guidance. When sampling from the base SBDD model trained on crystallographic data, the generative process will tend to be biased towards a small region of chemical space (orange). Ensemble guidance allows us to guide the structure-conditioned denoising process to generate more diverse and bio-active ligands using $\nabla \log p_{\text{ZINC}}$ (blue).

ZINC ZINC is a database of over 230 million commercially available compounds for virtual screening (Irwin & Shoichet, 2005). A random subset of 250,000 molecules is commonly used in machine learning (Gómez-Bombarelli et al., 2018).

2.2. Ensemble guidance allows for a 'mixture-of-chemists' in molecule design

We first consider a conditional score model trained only on protein-ligand complexes from the PDB distribution, $p_{\text{PDB}}(x_t|y)$, this model is functionally equivalent to to existing models like DiffSBDD (Schneuing et al., 2022). If one was interested in biasing this model to produce more bioactive compounds, we would first pretrain out model to generate random molecules from ZINC. However, this is likely to result in 'forgetting' novel scaffolds seen in the pretraining dataset and the final outputs will again be heavily biased towards the kinds of non-druglike ligands observed in the PDB.

Instead, we propose an alternative approach that does that lead to loss of information and is fully controllable at inference time. The simplest case was originally inspired by classifier-free guidance (Ho & Salimans, 2022), but can be viewed as a general 'mixture-of-chemists' approach that can be expended in number of ways. Namely, we augment the diffusion model trained on the PDB distribution p_{PDB} by training a new model on ZINC250k. While generation is unconditional, the molecules this model generates are highly diverse and drug-like. We then use this model to guide the structure-conditioned generation model:

$$\nabla \log p(x_t|y) = \gamma \underbrace{\nabla \log p_{\text{PDB}}(x_t|y)}_{\text{conditional but small}} + (1-\gamma) \underbrace{\nabla \log p_{\text{ZINC}}(x_t)}_{\text{unconditional but very large}}$$

where γ is a guidance term determining the weighting, which we assume to be constant here but could evolve during generation. While we implemented the simplest example as a proof of concept, ensemble guidance can be viewed as a general 'mixture-of-chemists' technique in molecule design that could be further extended (see Section 4). See Appendix A for details on implemenation.

3. Results

3.1. Interpolation between multiple chemists

We sample from our 'mixture-of-chemists' models using ensemble guidance ($\gamma \in [0-1]$). For now, we assume γ is a constant that does not evolve during training. We first measure the distribution of QED values as we vary γ to verify that we can effectively interpolate between the two learnt distributions (Fig 3A). We also found that PDB generated molecules were highly biased towards a high oxygen content (see Section 3.2), and that even minimal amount of ensemble guidance brought the atom frequencies more in line with ZINC (Fig 3B). To show the large differences in the learnt distributions between the two expert models, we perform t-SNE dimenionality reduction of the fingerprints from the molecules generated by each model (Fig 3C)

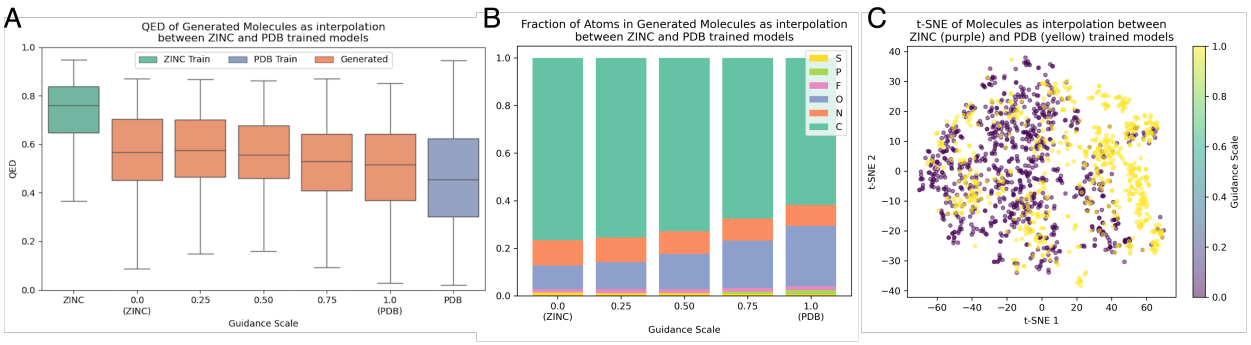

Figure 3. (A) Distribution of QED values for the ZINC and PDB datasets and molecules generated by varying ensemble guidance strength γ. (B) Impact of ensemble guidance on the frequency of atom types. (C) t-SNE plot showing the difference in learned distributions between the PDB and ZINC model.

3.2. Ensemble sampling guides generated molecules towards more bioactive substructures

We observed that models trained on the PDB tended to mode collapse on structures that were highly aliphatic (i.e. containing lots of branches rather than rings) and contained a large number of polar hydroxyl groups, which would suggest that these molecules are not suitably drug-like due to low membrane permeability. We perform SMARTS-based substructure matching between commonly observed substructures in ZINC and the PDB to determine their prevalence.

Figure 4 shows that there is a significantly higher prevalence of hydroxyl, glycosyl and phosphate groups both from the models trained on the PDB data as well as in the PDB training data itself. In all cases, we find that ensemble guidance with $\gamma = 0.5$ significantly improves the composition of substructures of the generative models and brings it in line with ZINC. This is quite striking in the case of phosphates, here ensemble guidance of only $\gamma = 0.5$, reduced the number of molecules with phosphate groups to negligible quantities. This highlights the ability of ensemble guidance to pull samples towards distributions of molecules with more favourable properties, overcoming limitations imposed by exclusive use of crystallographic data.

4. Discussion

Limitations So far our study has only examined the effect of guidance on the intrinsic physicochemical properties of generated molecules. It is likely that metrics assessing the quality of the generated poses, such as those proposed by Harris et al. (2023a), would suffer as the 3D-conditioned model is pushed to generate molecules outside of its training distribution. However, this does not necessarily preclude the ability of the model to generate out of distribution binders *per se* as these models already produce poses of dubious quality which can, to some extent, be rescued by traditional physics-based docking techniques. Furthermore, our use of coarse C_α pocket representations is likely to exacerbate this, though we expect embeddings from pre-trained protein structure encoders can be useful supplements (Zhang et al., 2022). However, taking 3DSBDD models to generate low-quality poses, we still believe we highlight meaningful limitations in the datasets used by the community and that ensemble guidance serves a useful purpose increasing the diversity and drug-likeness of designs. Further work will examine the influence of ensemble guidance on pose generation in further detail.

Future Work While we have demonstrated preliminary results suggesting the efficacy and viability of ensemble guidance, we propose that scaling ensemble guidance through composing additional generative models can enable the incorporation of much greater quantities of data in a controllable manner. We believe this effect will be synergistic, as each model can be specialised to incorporate favourable signal present in datasets of different scale and quality. For example, CrossDocked can be seen as a silver-standard dataset that can provide additional pose signal through a structure-conditioned model though adds limited chemical diversity, and an unconditional Enamine library can provide greater diversity with limited pose signal in a manner similar to ZINC.

$$\nabla \log p(x_t|y) = \alpha \underbrace{\nabla \log p_{\text{PDB}}(x_t|y)}_{\text{gold standard but small}} + \beta \underbrace{\nabla \log p_{\text{CrossDocked}}(x_t|y)}_{\text{silver standard but large}} + \gamma \underbrace{\nabla \log p_{\text{ZINC}}(x_t)}_{\text{unconditional but very large}}$$

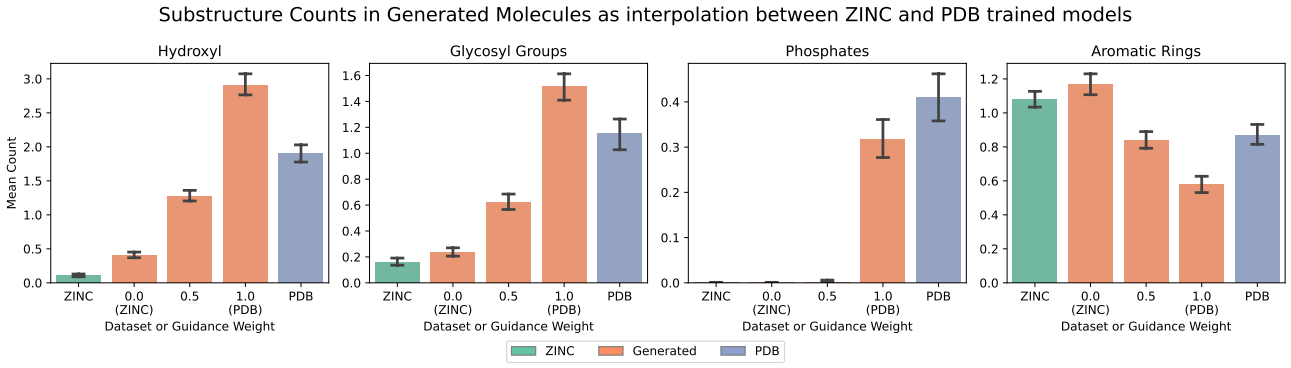

Figure 4. Increased guidance weight produces designs with interpolates substructure and motif occurrence between the different training distributions for hydroxyl groups, glycosyl groups, phosphates and aromatic rings. X-axis indicates either the dataset or the guidance weight that produced those molecules (effective training distribution placed below in brackets when appropriate).

Furthermore, explorations of more granular contributions to the guidance process could enable further gains, by allowing certain models to contribute more to different components of the generative process, such as atom type and bond placement.

5. Conclusion

In this work, we have demonstrated the drawbacks of exclusively relying on the PDB for training generative models for Structure-based Drug Design. Namely, the highly biased nature and limited diversity of the dataset presents a significant limitation to successful real-world application of these models in drug discovery and development campaigns.

To help address this, we introduce *ensemble guidance*, where we employ an ensemble of diffusion models trained on a variety of datasets to guide a structure-conditioned SBDD model to generate more diverse and bioactive compounds. Future work will focus on scaling up this approach to train models on millions of docked poses and billions of compounds from ultra-large virtual screening libraries.

References

Blundell, T. L. Structure-based drug design. *Nature*, 384 (6604):23, 1996.

Brocidiacono, M., Popov, K. I., Koes, D. R., and Tropsha, A. Plantain: Diffusion-inspired pose score minimization for fast and accurate molecular docking. *ArXiv*, 2023.

Drotár, P., Jamasb, A. R., Day, B., Cangea, C., and Liò, P. Structure-aware generation of drug-like molecules. *arXiv preprint arXiv: 2111.04107*, 2021.

Francoeur, P. G., Masuda, T., Sunseri, J., Jia, A., Iovanisci, R. B., Snyder, I., and Koes, D. R. Three-dimensional convolutional neural networks and a cross-docked data set for structure-based drug design. *Journal of Chemical Information and Modeling*, 60(9):4200–4215, August 2020. ISSN 1549-960X. doi: 10.1021/acs.jcim.0c00411. URL http://dx.doi.org/10.1021/acs.jcim.0c00411.

Gao, W. and Coley, C. W. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 60(12):5714–5723, 2020.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. *arXiv preprint arXiv:2303.03543*, 2023.

Halgren, T. A. Mmff vi. mmff94s option for energy minimization studies. *Journal of computational chemistry*, 20 (7):720–729, 1999.

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

Harris, C., Didi, K., Schneuing, A., Du, Y., Jamasb, A. R., Bronstein, M. M., Correia, B., Lio, P., and Blundell, T. L. Flexible small-molecule design and optimization with equivariant diffusion models. In *ICLR 2023-Machine Learning for Drug Discovery workshop*, 2023b.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

Igashov, I., Stärk, H., Vignac, C., Satorras, V. G., Frossard, P., Welling, M., Bronstein, M., and Correia, B. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

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

Jing, B., Eismann, S., Suriana, P., Townshend, R. J., and Dror, R. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

Jing, B., Eismann, S., Soni, P. N., and Dror, R. O. Equivariant graph neural networks for 3d macromolecular structure. *arXiv preprint arXiv:2106.03843*, 2021.

Keserű, G. M. and Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug discovery today*, 11(15-16): 741–748, 2006.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

Konc, J. and Janežič, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9):1160–1168, March 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btq100. URL http://dx.doi.org/10.1093/bioinformatics/btq100.

Peng, X., Luo, S., Guan, J., Xie, Q., Peng, J., and Ma, J. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International Conference on Machine Learning*, pp. 17644–17655. PMLR, 2022.

Polishchuk, P. G., Madzhidov, T. I., and Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *Journal of computer-aided molecular design*, 27:675–679, 2013.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Schneuing, A., Du, Y., Harris, C., Jamasb, A., Igashov, I., Du, W., Blundell, T., Lió, P., Gomes, C., Welling, M., et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Torge, J., Harris, C., Mathis, S. V., and Lio, P. Diffhopp: A graph diffusion model for novel drug design via scaffold hopping. *arXiv preprint arXiv:2308.07416*, 2023.

Zhang, Z., Xu, M., Jamasb, A. R., Chenthamarakshan, V., Lozano, A., Das, P., and Tang, J. Protein representation learning by geometric structure pretraining. *International Conference On Learning Representations*, 2022. doi: 10.48550/arXiv.2203.06125.

Zhang, Z., Zheng, S., Min, Y., and Liu, Q. Molecule generation for target protein binding with structural motifs. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Rq13idF0F73.

A. Implementation

Base models We train unconditional equivariant diffusion models as in Hoogeboom et al. (2022) and conditional SBDD diffusion models in Schneuing et al. (2022). All models use a Geometric Vector Preceptor (GVP) (Jing et al., 2020; 2021) as the denoiser network, as this was found to perform well in previous work (Torge et al., 2023). The conditional model was trained on only C_α-level granularity due to computational constraints. All models contain five layers with 128 and 64 scalar and vector features respectively. All models are trained on a single NVIDIA A100 GPU for seven days with a learning rate of 0.0001 using the Adam optimizer (Kingma & Ba, 2014).

Using non-3D chemical libaries In the case of ZINC 250K, molecules are given as SMILES without 3D information. Hence, we initialise conformers (without protein context) using the MMFF forcefield (Halgren, 1999) implementation in RDKIT to train our unconditional model. This is sufficient for our purposes as this data is to be used to guide samples towards highly diverse scaffolds/chemotypes, while the sample is guided to be a valid pose exhibiting high pocket-complementarity by the model trained on $p_{\text{PDB}}(x_t|y)$.

Chemoinformatics analysis We use ECFP fingerprints (Rogers & Hahn, 2010) with a maximum path length of 2 and 2,048 bits and perform Butina clustering at 50% Tanimoto similarity.