STRUCTURE-BASED DRUG DESIGN WITH EQUIVARIANT DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structure-based drug design (SBDD) aims to design small-molecule ligands that bind with high affinity and specificity to pre-determined protein targets. Traditional SBDD pipelines start with large-scale docking of compound libraries from public databases, thus limiting the exploration of chemical space to existent previously studied regions. Recent machine learning methods approached this problem using an atom-by-atom generation approach, which is computationally expensive. In this paper, we formulate SBDD as a 3D-conditional generation problem and present DiffSBDD, an SE(3)-equivariant 3D-conditional diffusion model that generates novel ligands conditioned on protein pockets. Furthermore, we curate a new dataset of experimentally determined binding complex data from Binding MOAD to provide a realistic binding scenario that complements the synthetic CrossDocked dataset. Comprehensive in silico experiments demonstrate the efficiency of DiffSBDD in generating novel and diverse drug-like ligands that engage protein pockets with high binding energies as predicted by in silico docking.

1 INTRODUCTION

The rational design of molecular compounds to act as drugs remains an oustanding challenge in biopharmaceutical research. Towards supporting such efforts, structure-based drug design (SBDD) aims to generate small-molecule ligands that bind to a specific 3D protein structure with high affinity and specificity (Anderson, 2003). However, SBDD remains very challenging and with important limitations. A traditional SBDD campaign starts with the identification and validation of a target of interest and its subsequent structural characterisation using experimental structural determination methods. The first step in this process is the identification of the binding pocket; a cavity in which ligands may bind the target to elicit the desired therapeutic effect. This can be achieved via experimental means or a plethora of computational approaches (Pérot et al., 2010). Once a binding site is identified, the goal is to discover lead compounds that exhibit the desired biological activity. Importantly, to transition from leads to promising candidates the compounds need to be evaluated regarding other drug development constraints that are also hard to predict (toxicity, absorption, etc.).

Traditionally, SBDD is handled either by high-throughput experimental or virtual screening (Lyne, 2002; Shoichet, 2004) of large chemical databases. Not only is this expensive and time consuming but it also limits the exploration of chemical space to the historical knowledge of previously studied molecules, with a further emphasis usually placed on commercial availability (Irwin & Shoichet, 2005). Moreover, the optimization of initial lead molecules is often a biased process, with heavy reliance on human intuition (Ferreira et al., 2015).

Recent advances in geometric deep learning, especially in modeling geometric structures of biomolecules (Bronstein et al., 2021; Atz et al., 2021), provide a promising direction for structure-based drug design (Gaudelet et al., 2021). Even though utilizing deep learning as surrogate docking models has achieved remarkable progress (Lu et al., 2022; Stärk et al., 2022), deep learning-based design of ligands that bind to target proteins is still an open problem. Early attempts have been made to represent molecules as atomic density maps, and variational auto-encoders were utilized to generate new atomic density maps corresponding to novel molecules (Ragoza et al., 2022). However, it is nontrivial to map atomic density maps back to molecules, necessitating a subsequent atom-fitting stage. Follow-up work addressed this limitation by representing molecules as 3D graphs with atomic coordinates and types which circumvents the unnecessary post-processing

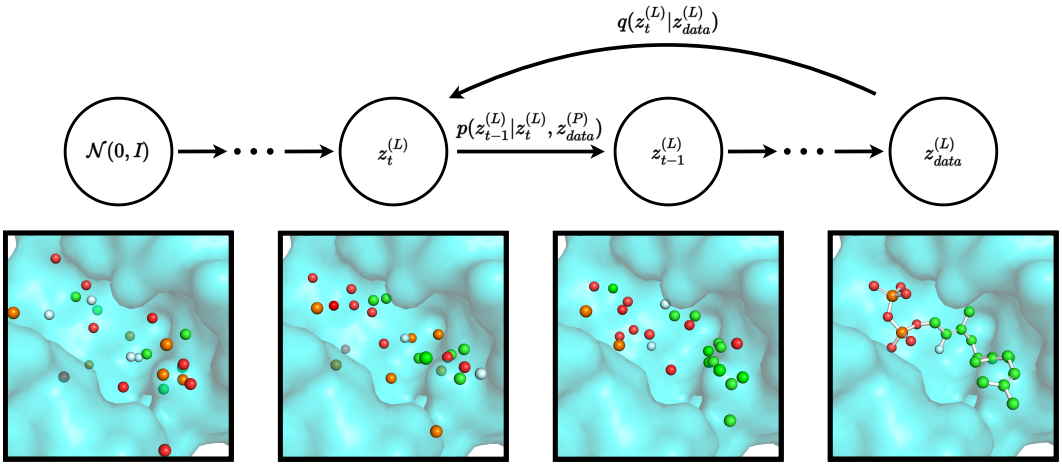

Figure 1: DiffSBDD in the protein-conditioned scenario. We first simulate the forward diffusion process q to gain a trajectory of progressively noised samples over T timesteps. We then train a model p_θ to reverse or denoise this process that is conditional on the target structure. Once trained, we are able to sample new drug candidates from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Both atom features and coordinates are diffused throughout the process. Ligands ($z^{(L)}$) are represented as fully-connected graphs during the diffusion process (edges not shown for clarity) and covalent bonds are added to the resultant point cloud at the end of generation. The protein ($z^{(P)}$) is represented as a graph but is shown as a surface here for clarity.

steps. Li et al. (2021) proposed an autoregressive generative model to sample ligands given the protein pocket as a conditioning constraint. Peng et al. (2022) improved this method by using an $E(3)$-equivariant graph neural network which respects rotation and translation symmetries in 3D space. Similarly, Drotár et al. (2021); Liu et al. (2022) used autoregressive models to generate atoms sequentially and incorporate angles during the generation process. Li et al. (2021) formulated the generation process as a reinforcement learning problem and connected the generator with Monte Carlo Tree Search for protein pocket-conditioned ligand generation. However, the main premise of sequential generation methods may not hold in real scenarios, since there is no ordering of the generation process and, as a result, the global context of the generated ligands may be lost. In addition, sequential methods pose more computational complexities that make the model inference inefficient (Luo et al., 2021; Peng et al., 2022).

An alternative is a one-shot generation strategy that samples the atomic coordinates and types of all the atoms at once (Du et al., 2022b). In this work, we develop an equivariant diffusion model for structure-based drug design (DiffSBDD) which, to the best of our knowledge, is the first of its kind. Specifically, we formulate SBDD as a 3D-conditioned generation problem where we aim to generate diverse ligands with high binding affinity for specific protein targets. We propose an $SE(3)$-equivariant 3D-conditional diffusion model that respects translation, rotation, and permutation equivariance. We introduce two strategies, *protein-conditioned generation* and *ligand-inpainting generation* producing new ligands conditioned on protein pockets. Specifically, protein-conditioned generation considers the protein as a fixed context, while ligand-inpainting models the joint distribution of the protein-ligand complex and new ligands are inpainted at inference time. We also demonstrate that our model can be used for out-of-the-box for molecular optimization. We further curate an experimentally determined binding dataset derived from Binding MOAD (Hu et al., 2005), which supplements the commonly used synthetic CrossDocked (Francoeur et al., 2020) dataset to validate our model performance under realistic binding scenarios. The experimental results demonstrate that DiffSBDD is capable of generating novel, diverse and drug-like ligands with predicted high binding affinities to given protein pockets. The code is available at `https://anonymous.4open.science/r/DiffSBDD-AF75/`.

2 Background

Denoising Diffusion Probabilistic Models Denoising diffusion probabilistic models (DDPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) are a class of generative models in-

spired by non-equilibrium thermodynamics. Briefly, they define a Markovian chain of random diffusion steps by slowly adding noise to sample data and then learning the reverse of this process (typically via a neural network) to reconstruct data samples from noise.

In this work, we closely follow the framework developed by Hoogeboom et al. (2022). In our setting, data samples are atomic point clouds $\boldsymbol{z}_{\text{data}} = [\boldsymbol{x}, \boldsymbol{h}]$ with 3D geometric coordinates $\boldsymbol{x} \in \mathbb{R}^{N\times 3}$ and categorical features $\boldsymbol{h} \in \mathbb{R}^{N\times d}$, where N is the number of atoms. A fixed noise process

$$q(\boldsymbol{z}_t|\boldsymbol{z}_{\text{data}}) = \mathcal{N}(\boldsymbol{z}_t|\alpha_t \boldsymbol{z}_{\text{data}}, \sigma_t^2 \boldsymbol{I}) \tag{1}$$

adds noise to the data $\boldsymbol{z}_{\text{data}}$ and produces a latent noised representation $\boldsymbol{z}_t$ for $t = 0, \ldots, T$. α_t controls the signal-to-noise ratio $\text{SNR}(t) = \alpha_t^2/\sigma_t^2$ and follows either a learned or pre-defined schedule from $\alpha_0 \approx 1$ to $\alpha_T \approx 0$ (Kingma et al., 2021). We also choose a variance-preserving noising process (Song et al., 2020) with $\alpha_t = \sqrt{1-\sigma_t^2}$.

Since the noising process is Markovian, we can write the denoising transition from time step t to $s < t$ in closed form as

$$q(\boldsymbol{z_s}|\boldsymbol{z}_{\text{data}}, \boldsymbol{z}_t) = \mathcal{N}\Big(\boldsymbol{z}_s \Big| \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\boldsymbol{z}_t + \frac{\alpha_s \sigma_{t|s}^2}{\sigma_t^2}\boldsymbol{z}_{\text{data}}, \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}\boldsymbol{I}\Big) \tag{2}$$

with $\alpha_{t|s} = \frac{\alpha_t}{\alpha_s}$ and $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2\sigma_s^2$ following the notation of Hoogeboom et al. (2022). This true denoising process depends on the data sample $\boldsymbol{z}_{\text{data}}$, which is not available when using the model for generating new samples. Instead, a neural network ϕ_θ is used to approximate the sample $\hat{\boldsymbol{z}}_{\text{data}}$. More specifically, we can reparameterize Equation (1) as $\boldsymbol{z}_t = \alpha_t \boldsymbol{z}_{\text{data}} + \sigma_t \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and directly predict the Gaussian noise $\hat{\boldsymbol{\epsilon}}_\theta = \phi_\theta(\boldsymbol{z}_t, t)$. Thus, $\hat{\boldsymbol{z}}_{\text{data}}$ is simply given as $\hat{\boldsymbol{z}}_{\text{data}} = \frac{1}{\alpha_t}\boldsymbol{z}_t - \frac{\sigma_t}{\alpha_t}\hat{\boldsymbol{\epsilon}}_\theta$.

The neural network is trained to maximise the likelihood of observed data by optimising a variational lower bound on the data, which is equivalent to the simplified training objective (Ho et al., 2020; Kingma et al., 2021) $\mathcal{L}_{\text{train}} = \frac{1}{2}||\boldsymbol{\epsilon} - \phi_\theta(\boldsymbol{z}_t, t)||^2$ up to a scale factor (see Appendix A).

$E(n)$-equivariant Graph Neural Networks A function $f : \mathcal{X} \to \mathcal{Y}$ is said to be *equivariant* w.r.t. the group G if $f(g.\boldsymbol{x}) = g.f(\boldsymbol{x})$, where $g.$ denotes the action of the group element $g \in G$ on $\mathcal{X}$ and $\mathcal{Y}$ (Serre et al., 1977). Graph Neural Networks (GNNs) are learnable functions that process graph-structured data in a permutation-equivariant way, making them particularly useful for molecular systems where nodes do not have an intrinsic order. Permutation invariance means that $\text{GNN}(\mathbf{\Pi X}) = \mathbf{\Pi}\,\text{GNN}(\mathbf{X})$ where $\mathbf{\Pi} \in \Sigma_n$ is an $n \times n$ permutation matrix acting on the node feature matrix. Since the nodes of the molecular graph represent the 3D coordinates of atoms, we are interested in additional equivariance w.r.t. the Euclidean group $E(3)$ or rigid transformations. An $E(3)$-equivariant GNN (EGNN) satisfies $\text{EGNN}(\mathbf{\Pi X A} + \mathbf{b}) = \mathbf{\Pi}\,\text{EGNN}(\mathbf{X})\mathbf{A} + \mathbf{b}$ for an orthogonal 3×3 matrix $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$ and some translation vector $\mathbf{b}$ added row-wise.

In our case, since the nodes have both geometric atomic coordinates $\boldsymbol{x}$ as well as atomic type features $\boldsymbol{h}$, we can use a simple implementation of EGNN proposed by Satorras et al. (2021), in which the updates for features $\boldsymbol{h}$ and coordinates $\boldsymbol{x}$ of node i at layer l are computed as follows:

$$\boldsymbol{m}_{ij} = \phi_e(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}),\ \tilde{e}_{ij} = \phi_{\text{att}}(\boldsymbol{m}_{ij}) \tag{3}$$

$$\boldsymbol{h}_i^{l+1} = \phi_h(\boldsymbol{h}_i^l, \sum_{j\neq i} \tilde{e}_{ij}\boldsymbol{m}_{ij}) \tag{4}$$

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + \sum_{j\neq i} \frac{\boldsymbol{x}_i^l - \boldsymbol{x}_j^l}{d_{ij}+1}\phi_x(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}) \tag{5}$$

where ϕ_e, ϕ_{att}, ϕ_h and ϕ_h are learnable Multi-layer Perceptrons (MLPs) and d_{ij} and a_{ij} are the relative distances and edge features between nodes i and j respectively.

3 Equivariant Diffusion Models for SBDD

We utilize an equivariant DDPM to generate molecules and binding conformations jointly with respect to a specific protein target. We represent protein and ligand point clouds as fully-connected graphs that are further processed by EGNNs (Satorras et al., 2021). We consider two distinct approaches to 3D pocket conditioning: (1) a conditional DDPM that receives a fixed pocket representation as context in each denoising step, and (2) a model that approximates the joint distribution of ligand-pocket pairs combined with inpainting at inference time.

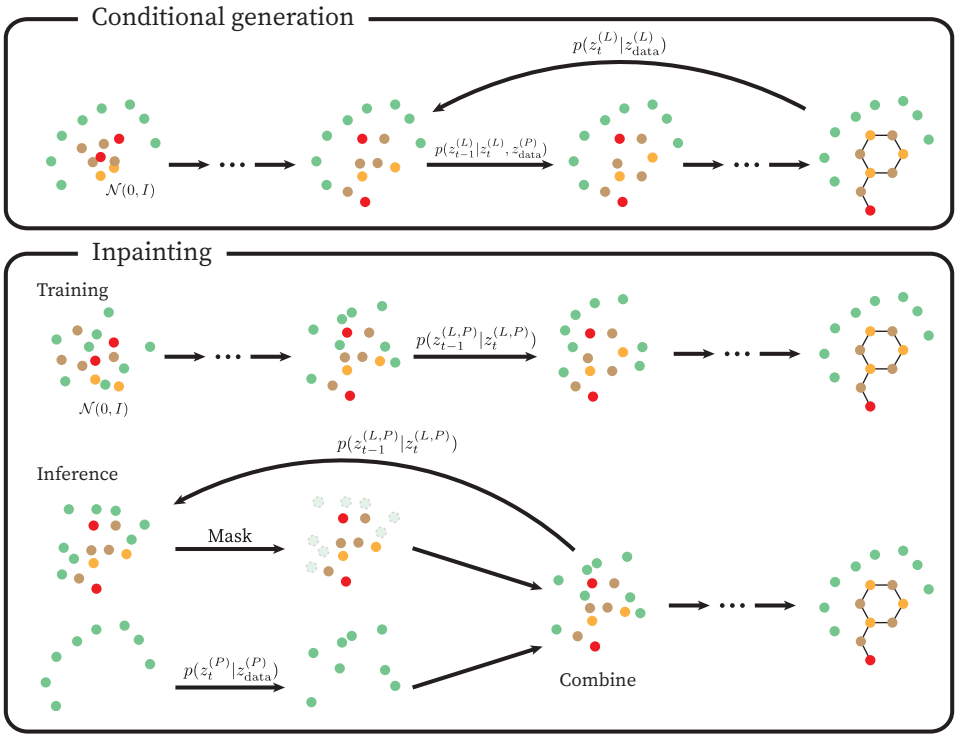

Figure 2: Comparison between the conditional generation and inpainting approaches. The conditional model learns to denoise molecules $\boldsymbol{z}^{(L)}$ in the fixed context of protein pockets $\boldsymbol{z}_{\text{data}}^{(P)}$. In the inpainting scenario, the model first learns to approximate the joint distribution of ligand and pocket nodes $\boldsymbol{z}_{\text{data}}^{(L,P)}$. For sampling, context is provided by combining the latent representation of the ligand with a forward diffused representation of the pocket in each denoising step.

3.1 Pocket-conditioned small molecule generation

In the conditional molecule generation setup, we provide fixed three-dimensional context in each step of the denoising process. To this end, we supplement the ligand node point cloud $\boldsymbol{z}_t^{(L)}$, denoted by superscript L, with protein pocket nodes $\boldsymbol{z}_{\text{data}}^{(P)}$, denoted by superscript P, that remain unchanged throughout the reverse diffusion process (Figure 2).

We parameterize the noise predictor $\hat{\boldsymbol{\epsilon}}_\theta = \phi_\theta(\boldsymbol{z}_t^{(L)}, \boldsymbol{z}_{\text{data}}^{(P)}, t)$ with an EGNN (Satorras et al., 2021; Hoogeboom et al., 2022). To process ligand and pocket nodes with a single GNN, atom types and residue types are first embedded in a joint node embedding space by separate learnable MLPs. We employ the same message-passing scheme outlined in Equations (3)-(5), however, following (Igashov et al., 2022) we do not update the coordinates of nodes that belong to the pocket to ensure the three-dimensional protein context remains fixed throughout the EGNN layers.

Equivariance In the probabilistic setting with 3D-conditioning, we would like to ensure $SE(3)$-equivariance in the following sense. This definition explicitly excludes reflections which are connected with chirality and can alter a biomolecule's properties.[1]:

- Evaluating the likelihood of a molecule $\boldsymbol{x}^{(L)} \in \mathbb{R}^{3\times N_L}$ given the three-dimensional representation of a protein pocket $\boldsymbol{x}^{(P)} \in \mathbb{R}^{3\times N_P}$ should not depend on global $SE(3)$-transformations of the system, i.e. $p(\boldsymbol{R}\boldsymbol{x}^{(L)}+\boldsymbol{t}|\boldsymbol{R}\boldsymbol{x}^{(P)}+\boldsymbol{t}) = p(\boldsymbol{x}^{(L)}|\boldsymbol{x}^{(P)})$ for orthogonal $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ with $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, $\det(\boldsymbol{R}) = 1$ and $\boldsymbol{t} \in \mathbb{R}^3$ added column-wise.
- At the same time, it should be possible to generate samples $\boldsymbol{x}^{(L)} \sim p(\boldsymbol{x}^{(L)}|\boldsymbol{x}^{(P)})$ from this conditional probability distribution so that equivalently transformed ligands $\boldsymbol{R}\boldsymbol{x}^{(L)}+\boldsymbol{t}$

[1]We transpose the node feature matrices hereafter so that the matrix multiplication resembles application of a group action. We also ignore node type features, which transform invariantly, for simpler notation.

are sampled with the same probability if the input pocket is rotated and translated and we sample from $p(\boldsymbol{R}\boldsymbol{x}^{(L)} + \boldsymbol{t} | \boldsymbol{R}\boldsymbol{x}^{(P)} + \boldsymbol{t})$.

Equivariance to the orthogonal group $O(3)$ (comprising rotations and reflections) is achieved because we model both prior and transition probabilities with isotropic Gaussians where the mean vector transforms equivariantly w.r.t. rotations of the context (see Hoogeboom et al. (2022) and Appendix E). Ensuring translation equivariance, however, is not as easy because the transition probabilities $p(\boldsymbol{z}_{t-1}|\boldsymbol{z}_t)$ are not inherently translation-equivariant. In order to circumvent this issue, we follow previous works (Köhler et al., 2020; Xu et al., 2022; Hoogeboom et al., 2022) by limiting the whole sampling process to a linear subspace where the center of mass (CoM) of the system is zero. In practice, this is achieved by subtracting the center of mass of the system before performing likelihood computations or denoising steps. Since equivariance of the transition probabilities depends on the parameterization of the noise predictor $\hat{\boldsymbol{\epsilon}}_\theta$, we can make the model sensitive to reflections with a simple additive term in the EGNN's coordinate update:

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + \sum_{j \neq i} \frac{\boldsymbol{x}_i^l - \boldsymbol{x}_j^l}{d_{ij}+1} \phi_x^d(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}) + \frac{(\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l) \times (\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)}{||(\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l) \times (\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)|| + 1} \phi_x^\times(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}), \tag{6}$$

using the cross product which changes sign under reflection. Here, $\bar{\boldsymbol{x}}^l$ denotes the center of mass of all nodes at layer l. $\phi_x^\times$ is an additional MLP. This modification is discussed in more detail in Appendix F.

3.2 Joint distribution with inpainting

As an extension to the conditional approach described above, we also present a ligand-*inpainting* approach. Originally introduced as a technique for completing masked parts of images (Song et al., 2020; Lugmayr et al., 2022), inpainting has been adopted in other domains, including biomolecular structures (Wang et al., 2022). Here, we extend this idea to three-dimensional point cloud data.

We first train an unconditional DDPM to approximate the joint distribution of ligand and pocket nodes $p(\boldsymbol{z}_{\text{data}}^{(L)}, \boldsymbol{z}_{\text{data}}^{(P)})$[2]. This allows us to sample new pairs without additional context. To condition on a target protein pocket, we then need to inject context into the sampling process by modifying the probabilistic transition steps. The combined latent representation $\boldsymbol{z}_{t-1}^{(L,P)}$ of protein pocket and ligand at diffusion step $t-1$ is assembled from a forward noised version of the pocket that is combined with ligand nodes predicted by the DDPM based on the previous latent representation at step t

$$\boldsymbol{z}_{t-1,\text{known}}^{(P)} \sim p(\boldsymbol{z}_{t-1}^{(P)} | \boldsymbol{z}_{\text{data}}^{(P)}) \tag{7}$$

$$\boldsymbol{z}_{t-1,\text{unknown}}^{(L,P)} \sim p_\theta(\boldsymbol{z}_{t-1}^{(L,P)} | \boldsymbol{z}_t^{(L,P)}) \tag{8}$$

$$\boldsymbol{z}_{t-1}^{(L,P)} = \left[\boldsymbol{z}_{t-1,\text{unknown}}^{(L)}, \boldsymbol{z}_{t-1,\text{known}}^{(P)}\right]. \tag{9}$$

In this manner, we traverse the Markov chain in reverse order from $t = T$ to $t = 0$, replacing the predicted pocket nodes with their forward noised counterparts in each step. Equation (8) conditions the generative process on the given protein pocket. Thanks to the noise schedule, which decreases the variance of the noising process to almost zero at $t = 0$ (Equation (1)), the final sample is guaranteed to contain an unperturbed representation of the protein pocket.

Since the model is trained to approximate the unconditional joint distribution of ligand-pocket pairs, the training procedure is identical to the unconditional molecule generation procedure developed by Hoogeboom et al. (2022) aside from the fully-connected neural networks that embed protein and ligand node features in a common space as described in Section 3.1. The conditioning on known protein pockets is entirely delegated to the sampling algorithm, which means this approach is not limited to ligand-inpainting but, in principle, allows us to mask and replace arbitrary parts of the ligand-pocket system without retraining.

Trippe et al. (2022) show that this simple *replacement method* inevitably introduces approximation error that can lead to inconsistent inpainted regions. In our experiments, we observe that the in-

[2]We use notations $\boldsymbol{z}^{(L,P)}$ and $[\boldsymbol{z}^{(L)}, \boldsymbol{z}^{(P)}]$ interchangeably to describe the combined system of ligand and pocket nodes.

Table 1: Evaluation of generated molecules for targets from the CrossDocked test set. * denotes that we re-evaluate the generated ligands provided by the authors. The inference times are taken from their papers. Note that these results have been produced with the E(3)-equivariant version of our model and will be updated.

	Vina Score (kcal/mol, ↓)	QED (↑)	SA (↑)	Lipinski (↑)	Diversity (↑)	Time (s, ↓)
Test set	-6.871 ± 2.32	0.476 ± 0.20	0.728 ± 0.14	4.340 ± 1.14	—	—
3D-SBDD (AR) (Luo et al., 2021)*	-5.888 ± 1.91	0.502 ± 0.17	0.675 ± 0.14	4.787 ± 0.51	0.742 ± 0.09	19659 ± 14704
Pocket2Mol (Peng et al., 2022)*	-7.058 ± 2.80	0.572 ± 0.16	0.752 ± 0.12	4.936 ± 0.27	0.735 ± 0.15	2504 ± 2207
DiffSBDD-cond (C_α)	-5.540 ± 1.57	0.460 ± 0.14	0.357 ± 0.09	4.821 ± 0.45	0.815 ± 0.06	324 ± 189
DiffSBDD-inpaint (C_α)	-5.735 ± 1.80	0.427 ± 0.15	0.343 ± 0.09	4.789 ± 0.49	0.807 ± 0.07	329 ± 177
DiffSBDD-cond	-6.584 ± 2.06	0.495 ± 0.15	0.336 ± 0.09	4.795 ± 0.49	0.730 ± 0.11	1634 ± 769

painting solution sometimes generates dislocated molecules that are not properly positioned in the target pocket. Trippe et al. (2022) propose to address this limitation with a particle filtering scheme that upweights more consistent samples in each denoising step. We, however, choose to adopt the conceptually simpler idea of *resampling* (Lugmayr et al., 2022), where each latent representation is repeatedly diffused back and forth before advancing to the next time step. This enables the model to harmonize its prediction for the unknown region and the noisy sample from the known region (Eq. (7)), which does not include any information about the generated part. We choose $r = 10$ resamplings per denoising step based on empirical results discussed in Appendix C.1.

Equivariance Similar desiderata as in the conditional case apply to the joint probability model, where we desire $SE(3)$-invariance that can be obtained from invariant priors via equivariant flows (Köhler et al., 2020). The main complications compared to the previous approach are the missing reference frame and impossibility of defining a valid translation-*invariant* prior noise distribution $p(\boldsymbol{z}_T)$ as such a distribution cannot integrate to one. Consequently, it is necessary to restrict the probabilistic model to a CoM-free subspace as described in previous works (Köhler et al., 2020; Xu et al., 2022; Hoogeboom et al., 2022). While the reverse diffusion process is defined for a CoM-free system, substituting the predicted pocket node coordinates with a new diffused version of the known pocket as described in Equations (7) - (9) can lead to non-zero CoM. To prevent this, we translate the known pocket representation so that its center of mass coincides with the predicted representation: $\tilde{\boldsymbol{x}}^{(P)}_{t-1,\text{known}} = \boldsymbol{x}^{(P)}_{t-1,\text{unknown}} - \boldsymbol{x}^{(P)}_{t-1,\text{known}}$ before creating the new combined representation $\boldsymbol{z}^{(L,P)}_{t-1} = [\boldsymbol{z}^{(L)}_{t-1,\text{unknown}}, \tilde{\boldsymbol{z}}^{(P)}_{t-1,\text{known}}]$ with $\tilde{\boldsymbol{z}}^{(P)}_{t-1,\text{known}} = [\tilde{\boldsymbol{x}}^{(P)}_{t-1,\text{known}}, \boldsymbol{h}^{(P)}_{t-1,\text{known}}]$.

4 Experiments

4.1 Datasets

CrossDocked We use the CrossDocked dataset (Francoeur et al., 2020) and follow the same filtering and splitting strategies as in previous work (Luo et al., 2021; Peng et al., 2022). This results in 100,000 high-quality protein-ligand pairs for the training set and 100 proteins for the test set. The split is done by 30% sequence identity using MMseqs2 (Steinegger & Söding, 2017).

Binding MOAD We also evaluate our method on experimentally determined protein-ligand complexes found in Binding MOAD (Hu et al., 2005) which are filtered and split based on the proteins' enzyme commission number as described in Appendix D. This results in 40,354 protein-ligand pairs for training and 130 pairs for testing.

4.2 Evaluation

For every experiment, we evaluated all combinations of all-atom and C_α level graphs with conditional and inpainting-based approaches respectively (with the exception of the all-atom inpainting approach due to computational limitations). Full details of model architecture and hyperparameters are given in Appendix C. We sampled 100 valid molecules[3] for each target pocket and removed all atoms that are not bonded to the largest connected fragment. Ligand sizes were sampled from

[3]Due to occasional processing issues the actual number of available molecules is slightly lower on average (see Appendix G.1).

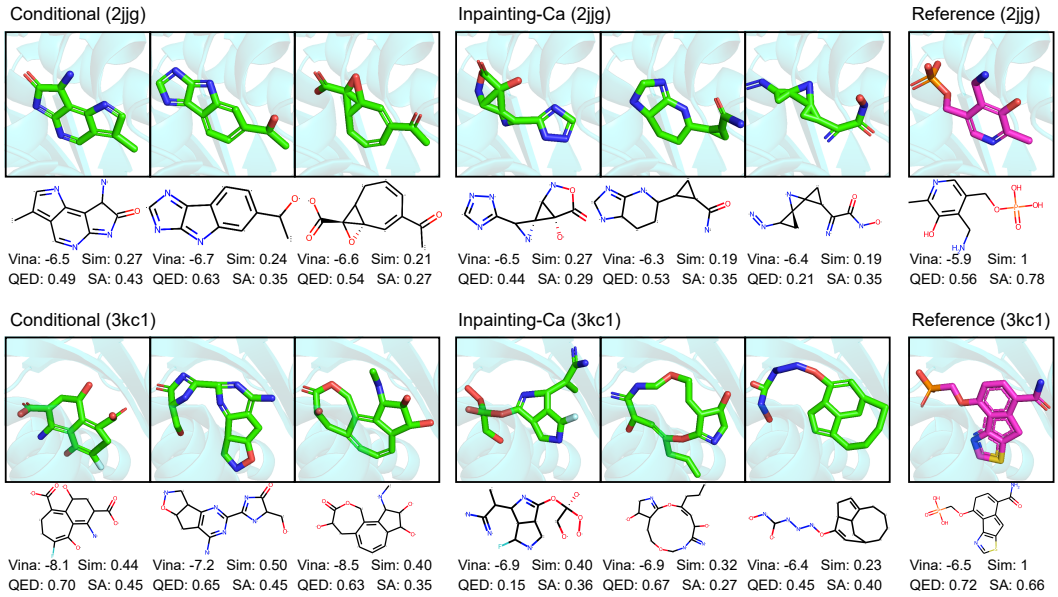

Figure 3: DiffSBDD models trained on CrossDocked and evaluated against a aminotransferase (top, PDB: 2jjg) and hydrolase (bottom, PDB: 3kc1). Conditional and inpainting approaches are compared (using all-atom and C_α level protein presentations respectively) and three high affinity molecules from each model are presented. 'Sim' is the Tanimoto similarity between the generated and reference ligand.

the training distribution as described in Appendix C. This procedure yields significantly smaller molecules than the reference ligands from the test set. We therefore increase the mean ligand size by 15 atoms to sample approximately equally sized molecules. This correction improves the observed docking scores which are highly correlated with the ligand size (see Figure C).

We employ widely-used metrics to assess the quality of our generated molecules (Peng et al., 2022; Li et al., 2021): (1) **Vina Score** is a physics-based estimation of binding affinity between small molecules and their target pocket; (2) **QED** is a simple quantitative estimation of drug-likeness combining several desirable molecular properties; (3) **SA** (synthetic accessibility) is a measure estimating the difficulty of synthesis; (4) **Lipinski** measures how many rules in the Lipinski rule of five (Lipinski et al., 2012), which is a loose rule of thumb to assess the drug-likeness of molecules, are satisfied; (5) **Diversity** is computed as the average pairwise dissimilarity (1 - *Tanimoto similarity*) between all generated molecules for each pocket; (6) **Inference Time** is the average time to sample 100 molecules for one pocket across all targets. All docking scores and chemical properties are calculated with QuickVina2 (Alhossary et al., 2015) and RDKit (Landrum et al., 2016).

4.3 Baselines

We compare with two recent deep learning methods for structure-based drug design. *3D-SBDD* (Luo et al., 2021) and Pocket2Mol (Peng et al., 2022) are auto-regressive schemes relying on graph representations of the protein pocket and previously placed atoms to predict probabilities based on which new atoms are added. *3D-SBDD* use heuristics to infer bonds from generated atomic point clouds while *Pocket2Mol* directly predicts them during the sequential generation process.

4.4 Results

CrossDocked Overall, the experimental results in Table 1 suggest that DiffSBDD can generate diverse small-molecule compounds with predicted high binding affinity, matching state-of-the-art performance. We do not see significant differences between the conditional model and the inpainting approach. The diversity score is arguably the most interesting, as this suggests our model is able to sample greater amounts of chemical space when compared to previous methods, while maintaining

Table 2: Evaluation of generated molecules for target pockets from the Binding MOAD test set.

	Vina Score (kcal/mol, ↓)	QED (↑)	SA (↑)	Lipinski (↑)	Diversity (↑)	Time (s, ↓)
Test set	-8.328 ± 2.05	0.602 ± 0.15	0.336 ± 0.08	4.838 ± 0.37	—	—
DiffSBDD-cond (C_α)	-6.281 ± 1.81	0.486 ± 0.17	0.313 ± 0.09	4.637 ± 0.63	0.730 ± 0.04	44.022 ± 8.98
DiffSBDD-inpaint (C_α)	-6.406 ± 5.13	0.512 ± 0.17	0.308 ± 0.09	4.681 ± 0.58	0.621 ± 0.16	98.439 ± 30.44
DiffSBDD-cond	-6.726 ± 1.60	0.470 ± 0.18	0.331 ± 0.08	4.666 ± 0.62	0.711 ± 0.08	194.860 ± 49.63

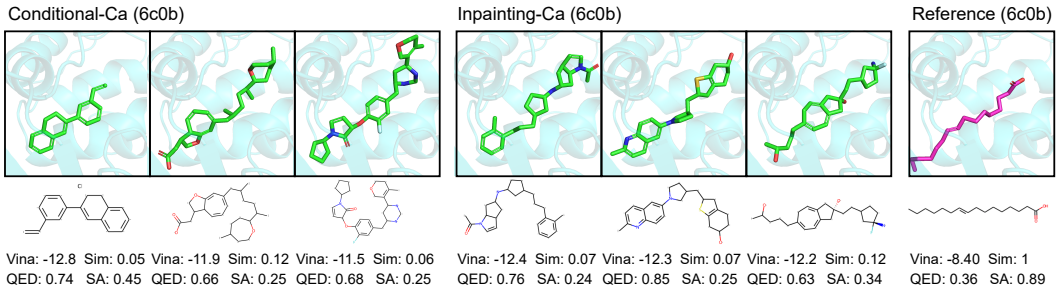

Figure 4: DiffSBDD models trained on Binding MOAD evaluated against a human receptor protein (PDB: 6c0b). Conditional and inpainting approaches are compared (C_α for both) and the three highest affinity molecules from each model are presented. Further details of the molecules shown here are explained in Appendix G.1

high binding performance, one of the most important requirements in early-stage, structure-based lead discovery. Specifically, DiffSBDD aims to generate ligands that bind to protein pockets and learn the probability density of ligands interacting with protein pockets. While it does not optimize for other molecular properties, such as QED and Lipinski, it generates molecules similar to the test set distributions. Only SA scores are significantly lower on average. While it is unclear why the model fails to approximate the distribution of synthetic accessibility scores successfully, simple techniques can be used for downstream optimisation of this property once promising candidates are found (Section 4.5). Generally, presenting the full atomic context to the model constrains the space of outputs considerably, leading to higher Vina scores but lower diversity compared to the C_α-only models. The all-atom model consistently beats C_α-based models on a per target basis (Appendix Figure 13).

A representative selection of molecules for two targets (*2jjg* and *3kc1*) are presented (Figure 3). This set is curated to be representative of our high scoring molecules, with both realistic and non-realistic motifs shown. It is noteworthy that the second molecule generated for 3kc1 has a similar tricyclic motif in the same pocket location as the reference ligand which was designed by traditional SBDD methods to maximise the hydrophobic interactions via shape complementarity of the ring system (Tsukada et al., 2010). However, a number of irregularities are present in even the highest scoring of generated molecules. For example, the high number of triangles in the molecules targeting 2jjg (from Inpainting-C_α) and the large rings for 3kc1 would prove difficult to synthesise. Random selections of generated molecules made by all methods evaluated are presented in Figure 11.

All docking scores reported in Table 1 are within one standard deviation of each other, which poses challenges for the discrimination of the best models. To verify successful pocket-conditioning, we therefore discuss the agreement of generated molecular conformations with poses after docking in Appendix G.5. This experiment showcases the success of our method to model protein-drug interactions at the atomic level and clearly highlights the benefits of the all-atom pocket representation.

Binding MOAD Results for the Binding MOAD dataset with experimentally determined binding complex data are reported in Table 2. 100 valid ligands have been generated for each of the 130 test pockets resulting in 13 000 molecules in total[4]. DiffSBDD generates highly diverse molecules but on average docking scores are lower than corresponding reference ligands from this dataset.

[4]The QuickVina score could not be computed for 49 ($\approx 0.4\%$) molecules from DiffSBDD-cond.

Generated molecules for a representative target are shown in Figure 4. The target (PDB: 6c0b) is a human receptor which is involved in microbial infection (Chen et al., 2018) and possibly tumor suppression (Ding et al., 2016). The reference molecule, a long fatty acid (see Figure 4) that aids receptor binding (Chen et al., 2018), has too high a number of rotatable bonds and low a number of hydrogen bond donors/acceptors to be considered a suitable drug (QED of 0.36). Our model however, generates drug-like (QED between 0.63-0.85) and suitably sized molecules by adding aromatic rings connected by a small number of rotatable bonds, which allows the molecules to adopt a complementary binding geometry and is entropically favourable (by reducing the degrees of freedom), a classic technique in medicinal chemistry (Ritchie & Macdonald, 2009). A random selection of generated molecules in presented in Figure 12.

4.5 Molecule Optimization

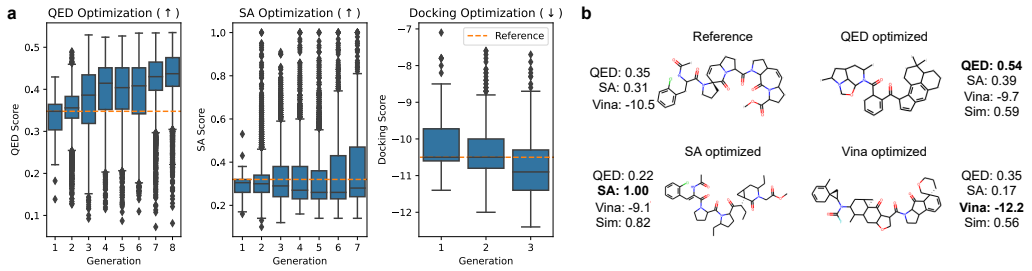

Figure 5: (a) Optimizing for various properties. (b) Examples of optimized molecules.

We use our model to optimise exciting candidate molecules, a common task in drug discovery called lead optimisation. This is when we take a compound found to have high binding affinity and optimise it for better 'drug-like' properties. We first noise the atom features and coordinates for t steps (where t is small) using the forward diffusion process. From this partially noised sample, we can then denoise the appropriate number of steps with the reverse process until $t = 0$. This allows us to sample new candidates of various properties whilst staying in the same region of active chemical space, assuming t is small (Appendix Figure G.4). This approach is inspired by (Luo et al., 2022) but note this does not allow for direct optimisation of specific properties, rather directed exploration around the local chemical space.

We extend this idea by combining the partial noising/denoising procedure with a simple evolutionary algorithm that selects for specific molecular properties. We find that our model performs well at this task out-of-the-box without any additional fine-tuning. As a showcase, we optimise a molecule in the test set targeting PDB:5NDU (a cancer therapeutic (Barone et al., 2020)), which has low SA and QED scores, 0.31 and 0.35 respectively, but high binding affinity. Over a number of rounds of optimisation, we can observe significant increases in QED (from 0.35 to mean of 0.43) whilst still maintaining high similarity to the original molecule (Figure 5a). We can also rescue the low synthetic accessibility score of the seed molecule by producing a battery of highly accessible molecules when selecting for SA (Figure 5b). Finally, we observe that we can performing significant optimsiation of binding affinity after only a view rounds of optimisation. Figure 5b shows 3 representative molecules with substantially optimised scores (QED, SA or Vina) whilst maintaining comparable binding affinity and globally similar structures. Full details are provided in Appendix 4.5.

5 Conclusion

In this work, we propose DiffSBDD, an $SE(3)$-equivariant 3D-conditional diffusion model for structure-based drug design. We demonstrate the effectiveness and efficiency of DiffSBDD in generating novel and diverse ligands with predicted high-affinity for given protein pockets on both a synthetic benchmark and a new dataset of experimentally determined protein-ligand complexes. We demonstrate that an inpainting-based approach can achieve competitive results to direct conditioning on a wide range of molecular metrics. Extending this more versatile strategy to an all atom pocket representation therefore holds promise to solve a variety of other structure-based drug design tasks, such as lead optimization or linker design, and binding site design without retraining.

REFERENCES

Keir Adams, Lagnajit Pattanaik, and Connor W Coley. Learning 3d representations of molecular chirality with invariance to bond rotations. *arXiv preprint arXiv:2110.04383*, 2021.

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.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

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

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

Matthias Barone, Matthias Müller, Slim Chiha, Jiang Ren, Dominik Albat, Arne Soicke, Stephan Dohmen, Marco Klein, Judith Bruns, Maarten van Dinther, et al. Designed nanomolar small-molecule inhibitors of ena/vasp evh1 interaction impair invasion and extravasation of breast cancer cells. *Proceedings of the National Academy of Sciences*, 117(47):29684–29690, 2020.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1): 1–11, 2022.

Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Peng Chen, Liang Tao, Tianyu Wang, Jie Zhang, Aina He, Kwok-ho Lam, Zheng Liu, Xi He, Kay Perry, Min Dong, et al. Structural basis for recognition of frizzled proteins by clostridium difficile toxin b. *Science*, 360(6389):664–669, 2018.

Lin-Can Ding, Xiao-Yu Huang, Fei-Fei Zheng, Jian Xie, Lin She, Yan Feng, Bo-Hua Su, Da-Li Zheng, and You-Guang Lu. Fzd2 inhibits the cell growth and migration of salivary adenoid cystic carcinomas. *Oncology Reports*, 35(2):1006–1012, 2016.

Pavol Drotár, Arian Rokkum Jamasb, Ben Day, Cătălina Cangea, and Pietro Liò. Structure-aware generation of drug-like molecules. *arXiv preprint arXiv:2111.04107*, 2021.

Weitao Du, He Zhang, Yuanqi Du, Qi Meng, Wei Chen, Nanning Zheng, Bin Shao, and Tie-Yan Liu. Se (3) equivariant graph neural networks with complete local frames. In *International Conference on Machine Learning*, pp. 5583–5608. PMLR, 2022a.

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

Yuanqi Du, Xian Liu, Nilay Shah, Shengchao Liu, Jieyu Zhang, and Bolei Zhou. Chemspace: Interpretable and interactive chemical space exploration. 2022c.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

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.

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.

Thomas Gaudelet, Ben Day, Arian R Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy B R Hayter, Richard Vickers, Charles Roberts, Jian Tang, David Roblin, Tom L Blundell, Michael M Bronstein, and Jake P Taylor-King. Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22(6), May 2021. doi: 10.1093/bib/bbab159. URL https://doi.org/10.1093/bib/bbab159.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

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

Lars Holdijk, Yuanqi Du, Ferry Hooft, Priyank Jaini, Bernd Ensing, and Max Welling. Path integral stochastic optimal control for sampling transition paths. *arXiv preprint arXiv:2207.02149*, 2022.

Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pp. 8867–8887. PMLR, 2022.

Liegi Hu, Mark L Benson, Richard D Smith, Michael G Lerner, and Heather A Carlson. Binding moad (mother of all databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

Ilia Igashov, Hannes Stärk, Clément Vignac, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274*, 2022.

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

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

Subha Kalyaanamoorthy and Yi-Ping Phoebe Chen. Structure-based drug design to augment hit discovery. *Drug discovery today*, 16(17-18):831–839, 2011.

Lawrence A Kelley, Stefans Mezulis, Christopher M Yates, Mark N Wass, and Michael JE Sternberg. The phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols*, 10(6): 845–858, 2015.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

Jonas Köhler, Leon Klein, and Frank Noé. Equivariant flows: exact likelihood generative learning for symmetric densities. In *International conference on machine learning*, pp. 5361–5370. PMLR, 2020.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Greg Landrum et al. Rdkit: Open-source cheminformatics software. 2016.

Kostiantyn Lapchevskyi, Benjamin Miller, Mario Geiger, and Tess Smidt. Euclidean neural networks (e3nn) v1. 0. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2020.

Yibo Li, Jianfeng Pei, and Luhua Lai. Structure-based de novo drug design using 3d deep generative models. *Chemical science*, 12(41):13664–13675, 2021.

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

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

Wei Lu, Qifeng Wu, Jixian Zhang, Jiahua Rao, Chengtao Li, and Shuangjia Zheng. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *bioRxiv*, 2022.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2837–2845, 2021.

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.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv*, 2022.

Paul D Lyne. Structure-based virtual screening: an overview. *Drug discovery today*, 7(20):1047–1055, 2002.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1): 1–14, 2011.

Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. *arXiv preprint arXiv:2205.07249*, 2022.

Stéphanie Pérot, Olivier Sperandio, Maria A Miteva, Anne-Claude Camproux, and Bruno O Villoutreix. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug discovery today*, 15(15-16):656–667, 2010.

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022.

Timothy J Ritchie and Simon JF Macdonald. The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug discovery today*, 14 (21-22):1011–1020, 2009.

Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Kristof T Schütt, Huziel E Sauceda, P-J Kindermans, Alexandre Tkatchenko, and K-R Müller. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

Jean-Pierre Serre et al. *Linear representations of finite groups*, volume 42. Springer, 1977.

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pp. 20503–20521. PMLR, 2022.

Martin Steinegger and Johannes Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, October 2017. doi: 10.1038/nbt.3988. URL https://doi.org/10.1038/nbt.3988.

Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Tomoharu Tsukada, Mizuki Takahashi, Toshiyasu Takemoto, Osamu Kanno, Takahiro Yamane, Sayako Kawamura, and Takahide Nishi. Structure-based drug design of tricyclic 8h-indeno [1, 2-d][1, 3] thiazoles as potent fbpase inhibitors. *Bioorganic & medicinal chemistry letters*, 20(3): 1004–1007, 2010.

Jue Wang, Sidney Lisanza, David Juergens, Doug Tischer, Joseph L Watson, Karla M Castro, Robert Ragotte, Amijai Saragovi, Lukas F Milles, Minkyung Baek, et al. Scaffolding protein functional sites using deep learning. *Science*, 377(6604):387–394, 2022.

Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.

Appendix for "Structure-based Drug Design with Equivariant Diffusion Models"

A Variational Lower Bound

To maximise the likelihood of our training data, we aim at optimising the variational lower bound (VLB) (Kingma et al., 2021; Hoogeboom et al., 2022)

$$-\log p(\boldsymbol{z}_{\text{data}}) \leq \underbrace{D_{\text{KL}}\big(q(\boldsymbol{z}_T|\boldsymbol{z}_{\text{data}})||p(\boldsymbol{z}_T)\big)}_{\text{prior loss } \mathcal{L}_{\text{prior}}} \underbrace{-\mathbb{E}_{q(\boldsymbol{z}_0|\boldsymbol{z}_{\text{data}})}\big[\log p(\boldsymbol{z}_{\text{data}}|\boldsymbol{z}_0)\big]}_{\text{reconstruction loss } \mathcal{L}_0} + \underbrace{\sum_{t=1}^{T} \mathcal{L}_t}_{\text{diffusion loss}} \tag{10}$$

with

$$\mathcal{L}_t = D_{\text{KL}}\big(q(\boldsymbol{z}_{t-1}|\boldsymbol{z}_{\text{data}}, \boldsymbol{z}_t)||p_\theta(\boldsymbol{z}_{t-1}|\hat{\boldsymbol{z}}_{\text{data}}, \boldsymbol{z}_t)\big) \tag{11}$$

$$= \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\Big[\frac{1}{2}\Big(\frac{\text{SNR}(t-1)}{\text{SNR}(t)} - 1\Big)||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_\theta||^2\Big] \tag{12}$$

during training. The prior loss should always be close to zero and can be computed exactly in closed form while the reconstruction loss must be estimated as described in Hoogeboom et al. (2022). In practice, however, we simply minimise the mean squared error $\mathcal{L}_{\text{train}} = \frac{1}{2}||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}||^2$ while randomly sampling time steps $t \sim \mathcal{U}(0, \ldots, T)$, which is equivalent up to a multiplicative factor.

B Note on Equivariance of the Conditional Model

The 3D-conditional model can achieve equivariance without the usual "subspace-trick". The coordinates of pocket nodes provide a reference frame for all samples that can be used to translate them to a unique location (e.g. such that the pocket is centered at the origin: $\sum_i \boldsymbol{x}_i^{(P)} = \mathbf{0}$). By doing this for all training data, translation equivariance becomes irrelevant and the CoM-free subspace approach obsolete. To evaluate the likelihood of translated samples at inference time, we can first subtract the pocket's center of mass from the whole system and compute the likelihood after this mapping. Similarly, for sampling molecules we can first generate a ligand in a CoM-free version of the pocket and move the whole system back to the original location of the pocket nodes to restore translation equivariance. As long as the mean of our Gaussian noise distribution $p(\boldsymbol{z}_t|\boldsymbol{z}_{\text{data}}^{(P)}) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{z}_{\text{data}}^{(P)}), \sigma^2\boldsymbol{I})$ depends equivariantly on the pocket node coordinates $\boldsymbol{x}^{(P)}$, $O(3)$-equivariance is satisfied as well (Appendix E). Since this change did not seem to affect the performance of the conditional model in our experiments, we decided to keep sampling in the linear subspace to ensure that the implementation is as similar as possible to the joint model, for which the subspace approach is necessary.

C Implementation Details

Molecule size As part of a sample's overall likelihood, we compute the empirical joint distribution of ligand and pocket nodes $p(N_L, N_P)$ observed in the training set and smooth it with a Gaussian filter ($\sigma = 1$). In the conditional generation scenario, we derive the distribution $p(N_L|N_P)$ and use it for likelihood computations.

For sampling, we can either fix molecule sizes manually or sample the number of ligand nodes from the same distribution given the number of nodes in the target pocket:

$$N_L \sim p(N_L|N_P). \tag{13}$$

For the experiments discussed in the main text, we increase the mean size of sampled molecules by 15 atoms to approximately match the sizes of molecules found in the test set. This modification makes the reported QuickVina scores more comparable as the *in silico* docking score is highly correlated with the molecular size, which is demonstrated in Figure C. With the correction, the average size of generated molecules is 26.5, 28.6, and 26 respectively for DiffSBDD-cond (C_α), DiffSBDD-inpaint (C_α) and DiffSBDD-cond (full atom). Test set molecules from the Binding MOAD data set are composed of 28 atoms on average.

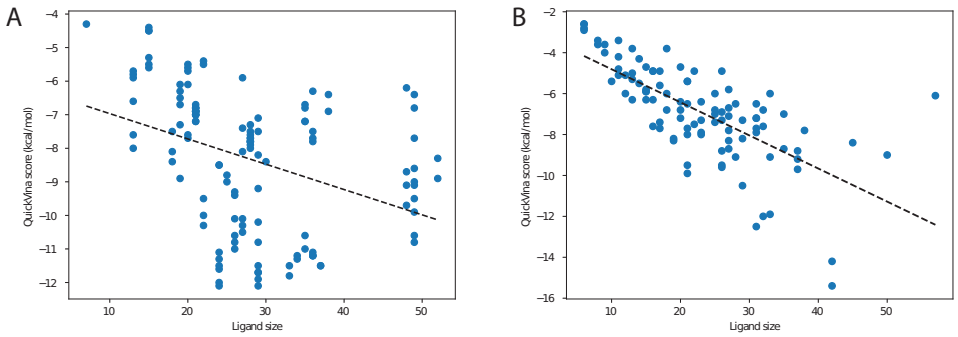

Figure 6: Correlation between ligand size and QuickVina score for reference molecules from the Binding MOAD (A) and CrossDocked (B) test sets.

Preprocessing All molecules are expressed as graphs. For the C_α only model the node features for the protein are set as the one hot encoding of the amino acid type. The full atom model uses the same one hot encoding of atom types for ligand and protein nodes. We refrain from adding a categorical feature for distinguishing between protein and ligand atoms in this case and continue using two separate MLPs for embedding the node features instead.

Noise schedule We use the pre-defined polynomial noise schedule introduced in (Hoogeboom et al., 2022):

$$\tilde{\alpha}_t = 1 - \left(\frac{t}{T}\right)^2, \quad t = 0, ..., T \tag{14}$$

Following (Nichol & Dhariwal, 2021; Hoogeboom et al., 2022), values of $\tilde{\alpha}^2_{t|s} = \left(\frac{\tilde{\alpha}_t}{\tilde{\alpha}_s}\right)^2$ are clipped between 0.001 and 1 for numerical stability near $t = T$, and $\tilde{\alpha}_t$ is recomputed as

$$\tilde{\alpha}_t = \prod_{\tau=0}^{t} \tilde{\alpha}_{\tau|\tau-1}. \tag{15}$$

A tiny offset $\epsilon = 10^{-5}$ is used to avoid numerical problems at $t = 0$ defining the final noise schedule:

$$\alpha_t^2 = (1 - 2\epsilon) \cdot \tilde{\alpha}_t^2 + \epsilon. \tag{16}$$

Feature scaling We scale the node type features $\boldsymbol{h}$ by a factor of 0.25 relative to the coordinates $\boldsymbol{x}$ which was empirically found to improve model perfomance in previous work (Hoogeboom et al., 2022).

Hyperparameters Hyperparameters for all presented models are summarized in Table 3. Training takes about $2.5\,\mathrm{h}/3.75\,\mathrm{h}$ (cond/inpaint) for Binding MOAD in the C_α scenario and $12.5\,\mathrm{h}$ with full atom pocket representation on a single NVIDIA V100. For CrossDocked, 100 training epochs take approximately $24\,\mathrm{h}$ on an NVIDIA V100 GPU in the C_α case and $96\,\mathrm{h}$ per 100 epochs on a single NVIDIA A100 GPU with all atom pocket representation.

Postprocessing For postprocessing of generated molecules, we use a similar procedure as in (Luo et al., 2021). Given a list of atom types and coordinates, bonds are first added using OpenBabel (O'Boyle et al., 2011). We then use RDKit to sanitise molecules, filter for the largest molecular fragment and finally remove steric clashes with 200 steps of force-field relaxation.

C.1 Effect of the number of resampling steps

In this section we discuss the number of resampling steps required for satisfactory results. To this end, we generated ligands for all test pockets with $r = 1$, $r = 5$, and $r = 10$ resampling steps, respectively. Because the resampling strategy slows down sampling approximately by a factor of r, we used the striding technique proposed by Nichol & Dhariwal (2021) and reduced the number of denoising steps proportionally to r. Nichol & Dhariwal (2021) showed that this approach reduces

Table 3: DiffSBDD hyperparameters.

	CrossDocked			Binding MOAD		
	Cond	Cond (C_α)	Inpaint (C_α)	Cond	Cond (C_α)	Inpaint (C_α)
No. layers	6	6	6	5	5	5
Joint embedding dim.	32	32	32	32	32	32
Hidden dim.	256	256	256	128	128	128
Learning rate	10^{-4}	10^{-4}	10^{-4}	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$	$5 \cdot 10^{-4}$
Weight decay	10^{-12}	10^{-12}	10^{-12}	10^{-12}	10^{-12}	10^{-12}
Diffusion steps	1000	1000	1000	500	500	500
Edges (ligand-ligand)	< 7 Å	fully connected	fully connected	fully connected	fully connected	fully connected
Edges (ligand-pocket)	< 7 Å	fully connected	fully connected	< 5 Å	< 8 Å	< 8 Å
Edges (pocket-pocket)	< 7 Å	fully connected	fully connected	< 5 Å	< 8 Å	< 8 Å
Epochs	1000	1000	1000	800	800	800

Table 4: Evaluation of generated molecules for target pockets from the Binding MOAD test set with the inpainting approach and C_α pocket representation for varying numbers of resampling steps r and denoising steps T.

r	T	Vina Score (kcal/mol, ↓)	QED (↑)	SA (↑)	Lipinski (↑)	Diversity (↑)	Time (s, ↓)
1	500	-5.601 ± 1.92	0.495 ± 0.12	0.358 ± 0.09	4.910 ± 0.31	0.850 ± 0.04	40.298 ± 13.52
5	100	-5.963 ± 1.93	0.541 ± 0.13	0.365 ± 0.09	4.946 ± 0.24	0.853 ± 0.05	45.074 ± 21.14
10	50	-6.080 ± 2.14	0.554 ± 0.13	0.367 ± 0.09	4.957 ± 0.21	0.855 ± 0.05	41.490 ± 14.32

the number of sampling steps significantly without sacrificing sample quality. In our case, it allows us to retain sampling speed while increasing the number of resampling steps.

To gauge the effect of resampling for molecule generation we show the distribution of RMSD values between the center of mass of reference molecules and generated molecules in Figure C.1. The unmodified replacement method ($r = 1$) produces molecules that are clearly farther away from the presumed pocket center than the conditional model. Increasing r moves the mean distance closer to the average displacement of molecules from the conditional method. This effect seems to saturate at $r = 10$ which is in line with the results obtained for images (Lugmayr et al., 2022).

Table 4 shows that neither the additional resampling steps nor the shortened denoising trajectory degrade the performance on the reported molecular metrics. The average docking scores even improve slightly which might reflect better positioning of generated ligands in the pockets prior to docking. The same model trained with $T = 500$ diffusion steps was used in all three cases. These experiments have been conducted without the adjustment of molecule sizes described in Section C.

D Binding MOAD Dataset

We curate a dataset of experimentally determined complexed protein-ligand structures from Binding MOAD (Hu et al., 2005). We keep pockets with valid[5] and moderately 'drug-like' ligands with QED score > 0.3. We further discard small molecules that contain atom types $\notin \{C, N, O, S, B, Br, Cl, P, I, F\}$ as well as binding pockets with non-standard amino acids. We define binding pockets as the set of residues that have any atom within 8 Å of any ligand atom. Ligand redundancy is reduced by randomly sampling at most 50 molecules with the same chemical component identifier (3-letter-code). After removing corrupted entries that could not be processed, 40 354 training pairs and 130 testing pairs remain. A validation set of size 246 is used to monitor estimated log-likelihoods during training. The split is made to ensure different sets do not contain proteins from the same Enzyme Commission Number (EC Number) main class.

E Proofs

In the following proofs we do not consider categorical node features $\boldsymbol{h}$ as only the positions $\boldsymbol{x}$ are subject to equivariance constraints. Furthermore, we do not distinguish between the zeroth latent

[5]as defined in `http://www.bindingmoad.org/`

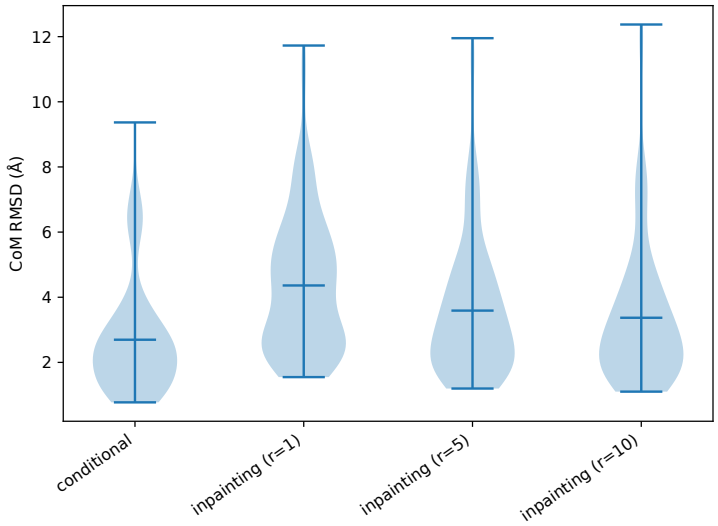

Figure 7: RMSD between reference molecules' center of mass and generated molecules' center of mass for the conditional model and inpaining model with varying numbers of resampling steps r. The pocket representation is C_α in all cases.

representation $\boldsymbol{x}_0$ and data domain representations $\boldsymbol{x}_{\text{data}}$ for ease of notation, and simply drop the subscripts.

E.1 $O(3)$-EQUIVARIANCE OF THE PRIOR PROBABILITY

The isotropic Gaussian prior $p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}) = \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{x}^{(\boldsymbol{P})}), \sigma^2\boldsymbol{I})$ is equivariant to rotations and reflections represented by an orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ as long as $\boldsymbol{\mu}(\boldsymbol{R}\boldsymbol{x}^{(P)}) = \boldsymbol{R}\boldsymbol{\mu}(\boldsymbol{x}^{(P)})$ because:

$$
\begin{aligned}
p(\boldsymbol{R}\boldsymbol{x}_T^{(L)}|\boldsymbol{R}\boldsymbol{x}^{(P)}) &= \frac{1}{\sqrt{(2\pi)^{N_L}\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}||\boldsymbol{R}\boldsymbol{x}_T^{(L)} - \boldsymbol{\mu}(\boldsymbol{R}\boldsymbol{x}^{(P)})||^2\Big) \\
&= \frac{1}{\sqrt{(2\pi)^{N_L}\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}||\boldsymbol{R}\boldsymbol{x}_T^{(L)} - \boldsymbol{R}\boldsymbol{\mu}(\boldsymbol{x}^{(P)})||^2\Big) \\
&= \frac{1}{\sqrt{(2\pi)^{N_L}\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}||\boldsymbol{R}\big(\boldsymbol{x}_T^{(L)} - \boldsymbol{\mu}(\boldsymbol{x}^{(P)})\big)||^2\Big) \\
&= \frac{1}{\sqrt{(2\pi)^{N_L}\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}||\boldsymbol{x}_T^{(L)} - \boldsymbol{\mu}(\boldsymbol{x}^{(P)})||^2\Big) \\
&= p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}).
\end{aligned}
$$

Here we used $||\boldsymbol{R}\boldsymbol{x}||_2 = ||\boldsymbol{x}||_2$ for orthogonal $\boldsymbol{R}$.

E.2 $O(3)$-EQUIVARIANCE OF THE TRANSITION PROBABILITIES

The denoising transition probabilities from time step t to $s < t$ are defined as isotropic normal distributions:

$$
p_\theta(\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{x}_t^{(L)}, \hat{\boldsymbol{x}}^{(L)}, \boldsymbol{x}^{(P)}) = \mathcal{N}(\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{\mu}_{t\to s}(\boldsymbol{x}_t^{(L)}, \hat{\boldsymbol{x}}^{(L)}, \boldsymbol{x}^{(P)}), \sigma_{t\to s}^2\boldsymbol{I}). \tag{17}
$$

Therefore, $p_\theta(\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{x}_t^{(L)}, \hat{\boldsymbol{x}}^{(L)}, \boldsymbol{x}^{(P)})$ is $O(3)$-equivariant by a similar argument to Section E.1 if $\boldsymbol{\mu}_{t\to s}$ is computed equivariantly from the three-dimensional context.

Recalling the definition of $\boldsymbol{\mu}_{t\to s} = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\boldsymbol{x}_t^{(L)} + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\hat{\boldsymbol{x}}^{(L)}$, we can prove its equivariance as follows:

$$\begin{aligned}
\boldsymbol{\mu}_{t\to s}(\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}) &= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\boldsymbol{R}\boldsymbol{x}_t^{(L)} + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\hat{\boldsymbol{x}}^{(L)}(\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}) \\
&= \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\boldsymbol{R}\boldsymbol{x}_t^{(L)} + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\boldsymbol{R}\hat{\boldsymbol{x}}^{(L)}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}) \quad \text{(equivariance of } \hat{\boldsymbol{x}}^{(L)}) \\
&= \boldsymbol{R}\Big(\frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\boldsymbol{x}_t^{(L)} + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\hat{\boldsymbol{x}}^{(L)}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)})\Big) \\
&= \boldsymbol{R}\boldsymbol{\mu}_{t\to s}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}),
\end{aligned}$$

where $\hat{\boldsymbol{x}}^{(L)}$ defined as $\hat{\boldsymbol{x}}^{(L)} = \frac{1}{\alpha_t}\boldsymbol{x}_t^{(L)} - \frac{\sigma_t}{\alpha_t}\hat{\boldsymbol{\epsilon}}$ is equivariant because:

$$\begin{aligned}
\hat{\boldsymbol{x}}^{(L)}(\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}) &= \frac{1}{\alpha_t}\boldsymbol{R}\boldsymbol{x}_t^{(L)} - \frac{\sigma_t}{\alpha_t}\hat{\boldsymbol{\epsilon}}(\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}, t) \\
&= \frac{1}{\alpha_t}\boldsymbol{R}\boldsymbol{x}_t^{(L)} - \frac{\sigma_t}{\alpha_t}\boldsymbol{R}\hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}, t) \quad (\hat{\boldsymbol{\epsilon}} \text{ predicted by equivariant neural network}) \\
&= \boldsymbol{R}\Big(\frac{1}{\alpha_t}\boldsymbol{x}_t^{(L)} - \frac{\sigma_t}{\alpha_t}\hat{\boldsymbol{\epsilon}}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}, t)\Big) \\
&= \boldsymbol{R}\hat{\boldsymbol{x}}^{(L)}(\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}).
\end{aligned}$$

E.3 $O(3)$-EQUIVARIANCE OF THE LEARNED LIKELIHOOD

Let $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ be an orthogonal matrix representing an element g from the general orthogonal group $O(3)$. We obtain the marginal probability density of the Markovian denoising process as follows

$$\begin{aligned}
p_\theta(\boldsymbol{x}_0^{(L)}|\boldsymbol{x}^{(P)}) &= \int p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}) p_\theta(\boldsymbol{x}_{0:T-1}^{(L)}|\boldsymbol{x}_T^{(L)}, \boldsymbol{x}^{(P)}) \mathrm{d}\boldsymbol{x}_{1:T} \\
&= \int p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}) \mathrm{d}\boldsymbol{x}_{1:T}
\end{aligned}$$

and the sample's likelihood is $O(3)$-equivariant:

$$\begin{aligned}
p_\theta(\boldsymbol{R}\boldsymbol{x}_0^{(L)}|\boldsymbol{R}\boldsymbol{x}^{(P)}) &= \int p(\boldsymbol{R}\boldsymbol{x}_T^{(L)}|\boldsymbol{R}\boldsymbol{x}^{(P)}) \prod_{t=1}^{T} p_\theta(\boldsymbol{R}\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}) \mathrm{d}\boldsymbol{x}_{1:T} \\
&= \int p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}) \prod_{t=1}^{T} p_\theta(\boldsymbol{R}\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{R}\boldsymbol{x}_t^{(L)}, \boldsymbol{R}\boldsymbol{x}^{(P)}) \mathrm{d}\boldsymbol{x}_{1:T} \quad \text{(equivariant prior)} \\
&= \int p(\boldsymbol{x}_T^{(L)}|\boldsymbol{x}^{(P)}) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}^{(L)}|\boldsymbol{x}_t^{(L)}, \boldsymbol{x}^{(P)}) \mathrm{d}\boldsymbol{x}_{1:T} \quad \text{(equivariant transition probabilities)} \\
&= p_\theta(\boldsymbol{x}_0^{(L)}|\boldsymbol{x}^{(P)})
\end{aligned}$$

F $SE(3)$-EQUIVARIANT GRAPH NEURAL NETWORK

Chiral molecules cannot be superimposed by combination of rotations and translations. Instead they are mirrored along a stereocenter, axis, or plane. As chirality can significantly alter a molecule's chemical properties, we use a variant of the $E(3)$-equivariant graph neural networks (Satorras et al., 2021) presented in Equations (3)-(5) that is sensitive to reflections and hence $SE(3)$-equivariant. We change the coordinate update equation, Equ. (5), of standard EGNNs in the following way

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + \sum_{j\neq i} \frac{\boldsymbol{x}_i^l - \boldsymbol{x}_j^l}{d_{ij}+1}\phi_x^d(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}) + \frac{(\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l)\times(\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)}{||(\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l)\times(\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)||+1}\phi_x^\times(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij}), \tag{18}$$

where $\bar{\boldsymbol{x}}^l$ denotes the center of mass of all nodes at layer l. This modification makes the EGNN layer sensitive to reflections while staying close to the original formalism. Since the resulting graph neural networks are only equivariant to the $SE(3)$ group, we will hereafter call them SEGNNs for short.

F.1 Discussion of Equivariance

Here we study how the suggested change in the coordinate update equation breaks reflection symmetry while preserving equivariance to rotations. Messages and scalar feature updates (Equations (3) and (4)) remain $E(3)$-invariant as in the original model and are therefore not considered in this section. We analyze transformations composed of a translation by $\boldsymbol{t} \in \mathbb{R}^3$ and a rotation/reflection by an orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ with $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$. The output at layer $l+1$ given the transformed input $\boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t}$ at layer l is calculated as:

$$\boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t} + \sum_{j\neq i} \frac{\boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t} - (\boldsymbol{R}\boldsymbol{x}_j^l + \boldsymbol{t})}{d_{ij}+1}\phi_x^d(\cdot) + \frac{(\boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t} - (\boldsymbol{R}\bar{\boldsymbol{x}}^l + \boldsymbol{t})) \times (\boldsymbol{R}\boldsymbol{x}_j^l + \boldsymbol{t} - (\boldsymbol{R}\bar{\boldsymbol{x}}^l + \boldsymbol{t}))}{Z_{ij}^{\times}+1}\phi_x^{\times}(\cdot) \tag{19}$$

$$= \boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t} + \sum_{j\neq i} \frac{\boldsymbol{R}(\boldsymbol{x}_i^l - \boldsymbol{x}_j^l)}{d_{ij}+1}\phi_x^d(\cdot) + \frac{(\boldsymbol{R}\boldsymbol{x}_i^l - \boldsymbol{R}\bar{\boldsymbol{x}}^l) \times (\boldsymbol{R}\boldsymbol{x}_j^l - \boldsymbol{R}\bar{\boldsymbol{x}}^l)}{Z_{ij}^{\times}+1}\phi_x^{\times}(\cdot) \tag{20}$$

$$= \boldsymbol{R}\boldsymbol{x}_i^l + \boldsymbol{t} + \sum_{j\neq i} \frac{\boldsymbol{R}(\boldsymbol{x}_i^l - \boldsymbol{x}_j^l)}{d_{ij}+1}\phi_x^d(\cdot) + \frac{\det(\boldsymbol{R})\boldsymbol{R}\big((\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l) \times (\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)\big)}{Z_{ij}^{\times}+1}\phi_x^{\times}(\cdot) \tag{21}$$

$$= \boldsymbol{R}\boldsymbol{x}_i^{l+1} + \boldsymbol{t} + \big(\det(\boldsymbol{R}) - 1\big)\sum_{j\neq i} \frac{\boldsymbol{R}\big((\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l) \times (\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)\big)}{Z_{ij}^{\times}+1}. \tag{22}$$

This result shows that the output coordinates are only equivariantly transformed if $\boldsymbol{R}$ is orientation preserving, i.e. $\det(\boldsymbol{R}) = 1$. If $\boldsymbol{R}$ is a reflection ($\det(\boldsymbol{R}) = -1$), coordinates will be updated with an additional summand that breaks the symmetry.

The learnable coefficients $\phi_x^d(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij})$ and $\phi_x^{\times}(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, d_{ij}^2, a_{ij})$ only depend on relative distances and are therefore $E(3)$-invariant. Their arguments are represented with the "$\cdot$" symbol for brevity. Likewise, the normalization factor $||(\boldsymbol{x}_i^l - \bar{\boldsymbol{x}}^l) \times (\boldsymbol{x}_j^l - \bar{\boldsymbol{x}}^l)||$ is abbreviated as $Z_{ij}^{\times}$. Already in the first line we used the fact that the mean transforms equivariantly. Furthermore, we use $\boldsymbol{R}\boldsymbol{a} \times \boldsymbol{R}\boldsymbol{b} = \det(\boldsymbol{R})\boldsymbol{R}(\boldsymbol{a} \times \boldsymbol{b})$ in the second step, which can be derived as follows:

$$\boldsymbol{x}^T(\boldsymbol{R}\boldsymbol{a} \times \boldsymbol{R}\boldsymbol{b}) = \det(\underbrace{[\boldsymbol{x}, \boldsymbol{R}\boldsymbol{a}, \boldsymbol{R}\boldsymbol{b}]}_{\in\mathbb{R}^{3\times 3}}) \tag{23}$$

$$= \det(\boldsymbol{R}[\boldsymbol{R}^T\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}]) \tag{24}$$

$$= \det(\boldsymbol{R})\det([\boldsymbol{R}^T\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}]) \tag{25}$$

$$= \det(\boldsymbol{R})\Big(\boldsymbol{x}^T\boldsymbol{R}(\boldsymbol{a} \times \boldsymbol{b})\Big) \tag{26}$$

$$= \boldsymbol{x}^T\Big(\det(\boldsymbol{R})\boldsymbol{R}(\boldsymbol{a} \times \boldsymbol{b})\Big) \tag{27}$$

The stated property of the cross product follows because this derivation is true for all $\boldsymbol{x} \in \mathbb{R}^3$.

F.2 EMPIRICAL RESULTS

To show the effectiveness of this architecture on a simple toy example, we repeat the classification experiment by Adams et al. (2021) who train neural networks to classify tetrahedral chiral centers as right-handed (*rectus*, 'R') or left-handed (*sinister*, 'S'). We closely follow their data split and experimental set-up and only replace the classifier with EGNN and SEGNNs, respectively. The results in Table 5 clearly demonstrate that the $SE(3)$-equivariant EGNN is capable of solving this task (without any hyperparameter optimization) whereas the $E(3)$-equivariant version does not do better than random guessing.

Table 5: Accuracy on the R/S classification task. Results from the first section are taken from (Adams et al., 2021) and included for reference. Mean and standard deviations are calculated based on three repetitions.

Model	R/S Accuracy (%)
ChIRo	98.5
SchNet	54.4
DimeNet++	65.7
SphereNet	98.2
EGNN	50.4
SEGNN	83.4

G EXTENDED RESULTS

G.1 ADDITIONAL EXPERIMENTAL DETAILS

The numbers of available molecules differ slightly between different methods due to computational issues or missing molecules in the available baseline sets. More precisely, on average 93.5, 92.8, and 98.3 molecules have been evaluated per pocket for DiffSBDD-cond, DiffSBDD-inpaint (C_α), and DiffSBDD-cond (C_α), respectively. For Pocket2Mol, 98.4 molecules are available per pocket. The set of 3D-SBDD molecules does not contain generated ligands for two test pockets. For the remaining 98 pockets, 89.9 molecules are available on average.

All Figures show molecules generated where the starting number of nodes equals the number of nodes in the reference ligands, with the exception of Figure 4, which employs the sampling strategy outlined in Appendix C.

G.2 ADDITIONAL MOLECULAR METRICS

In addition to the molecular properties discussed in Section 4 we assess the models' ability to produce novel and valid molecules using four simple metrics: validity, connectivity, uniqueness, and novelty. **Validity** measures the proportion of generated molecules that pass basic tests by RDKit–mostly ensuring correct valencies. **Connectivity** is the proportion of valid molecules that do not contain any disconnected fragments. We convert every valid and connected molecule from a graph into a canonical SMILES string representation, count the number unique occurrences in the set of generated molecules and compare those to the training set SMILES to compute **uniqueness** and **novelty** respectively.

Table 6 shows that only a small fraction of all generated molecules is invalid and must be discarded for downstream processing. The DiffSBDD models trained on CrossDocked with C_α pocket representation generate fragmented molecules about 50% of the time. Since we can simply select and process the largest fragments in these cases, low connectivity does not necessarily affect the efficiency of the generative process. Moreover, all models produce diverse sets of molecules unseen in the training set.

G.3 OCTANOL-WATER PARTITION COEFFICIENT

The octanol-water partition coefficient ($\log P$) is a measure of lipophilicity and is commonly reported for potential drug candidates (Wildman & Crippen, 1999). We summarize this property for our generated molecules in Table 7.

Table 6: Basic molecular metrics for generated small molecules given a C_α and full atom representation of the protein pocket.

Model	**Validity**	**Connectivity**	**Uniqueness**	**Novelty**
CrossDocked Training data	100%	100%	–	–
DiffSBDD-cond (C_α)	97.75%	48.02%	96.95%	100%
DiffSBDD-inpaint (C_α)	91.62%	51.38%	98.64%	100%
DiffSBDD-cond	93.23%	83.46%	97.46%	100%
Binding MOAD Training data	96.38%	100%	–	–
DiffSBDD-cond (C_α)	92.51%	52.13%	100.00%	100.00%
DiffSBDD-inpaint (C_α)	90.28%	73.19%	100.00%	100.00%
DiffSBDD-cond	95.39%	39.58%	100.00%	100.00%

Table 7: LogP values of generated molecules.

	CrossDocked	Binding MOAD
Test set	0.894 ± 2.73	0.456 ± 1.15
3D-SBDD (AR) (Luo et al., 2021)	0.273 ± 2.01	—
Pocket2Mol (Peng et al., 2022)	1.720 ± 1.97	—
DiffSBDD-cond (C_α)	-0.184 ± 1.01	0.110 ± 1.03
DiffSBDD-inpaint (C_α)	-0.519 ± 1.09	0.574 ± 1.39
DiffSBDD-cond	-0.328 ± 1.18	0.845 ± 1.61

G.4 Optimization

We demonstrate the effect the number of noising/denoising steps (t) has on various molecular properties in Figure G.4. We test all values of t at intervals of 10 steps and 200 molecules are sampled at every timestep. Note this does not allow for explicit optimization of any particular property unless combined with the evolutionary algorithm.

During the evolutionary algorithm, at the end of every generation the top 10 docking molecules are used to seed the next population. Every seed molecule is elaborated into 20 new candidates with a randomly chosen t between 10 and 150. To make the first population, we start with the single reference molecule and sample 200 new molecules with t chosen as above.

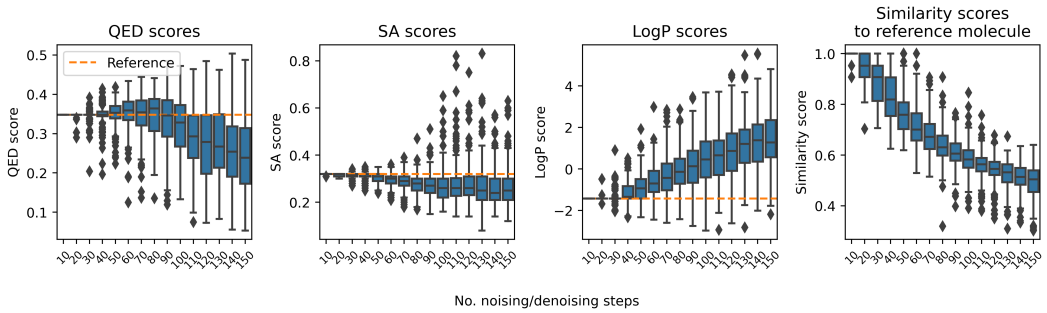

Figure 8: Effect of number of noising/denoising steps on molecule properties.

G.5 Agreement of generated and docked conformations

Here we discuss an alternative way of using QuickVina for assessing the quality of the conditional generation procedure besides its *in silico* docking score. We compare the generated raw conformations (before force-field relaxation) to the best scoring QuickVina docking pose and plot the

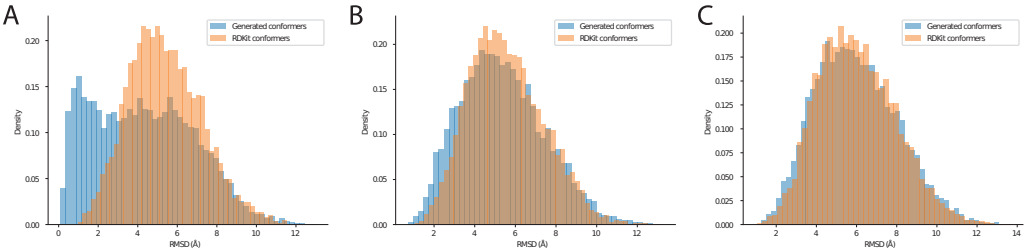

Figure 9: RMSD between original and docked conformations for the CrossDocked dataset. (A) DiffSBDD-cond, sample size 8804. (B) DiffSBDD-cond (C_α), sample size 9611. (C) DiffSBDD-inpaint (C_α), sample size 8641.

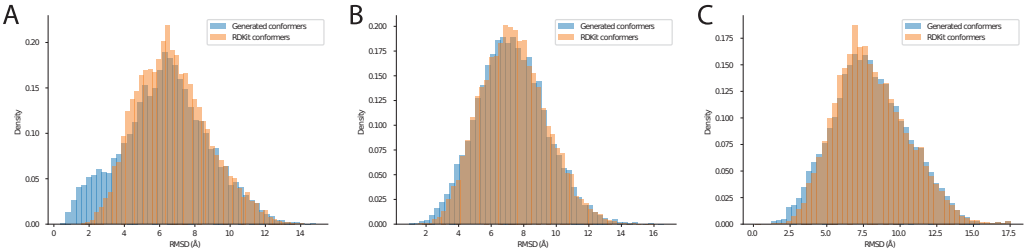

Figure 10: RMSD between original and docked conformations for the Binding MOAD dataset. (A) DiffSBDD-cond, sample size 12 520. (B) DiffSBDD-cond (C_α), sample size 12 641. (C) DiffSBDD-inpaint (C_α), sample size 12 902.

distribution of resulting RMSD values in Figures 9 and 10. As a baseline, the procedure is repeated for RDKit conformers of the same molecules with identical center of mass. For a large percentage of molecules generated by the all-atom CrossDocked model, QuickVina agrees with the predicted bound conformations, leaving them almost unchanged (RMSD below 2 Å). This demonstrates successful conditioning on the given protein pockets.

For the C_α-only models results are less convincing. They produce poses that only slightly improve upon conformers lacking pocket-context. Likely, this is caused by atomic clashes with the proteins' side chains that QuickVina needs to resolve.

G.6 Random generated molecules

Randomly selected molecules generated with our method and 3 baseline methods (LiGAN, SBDD-3D and Pocket2Mol) when trained with CrossDocked are presented in Figure 11. Randomly selected molecules generated by our method when trained with Binding MOAD are show in Figure 12.

G.7 Distribution of docking scores by target

We present extensive evaluation of the docking scores for our generated molecules in Figure 13. We evaluate all models trained with a given dataset first against all targets (Figure 13A+C) and 10 randomly chosen targets (Figure 13B+D). We note that the all-atom model trained using CrossDocked data outperforms all other methods. Unsurprisingly, model performance is highly target dependent, likely varying with properties like pocket geometry, size, charge, and hydrophbicity, which would affect the propensity of generating high affinity molecules.

H Related Work

Diffusion Models for Molecules Inspired by non-equilibrium thermodynamics, diffusion models have been proposed to learn data distributions by modeling a denoising (reverse diffusion) process and have achieved remarkable success in a variety of tasks such as image, audio synthesis and point

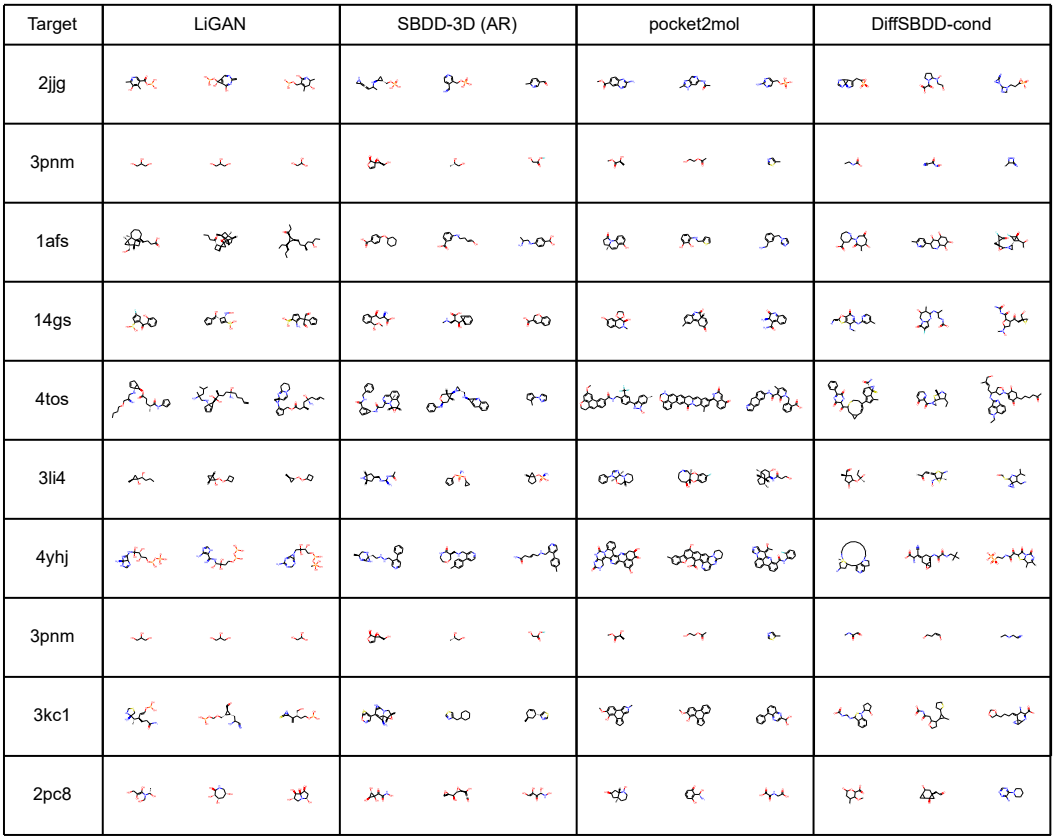

Figure 11: Generated molecules for 10 randomly chosen targets in the CrossDocked test set. For each target, 3 randomly selected generated molecules from 4 models are shown.

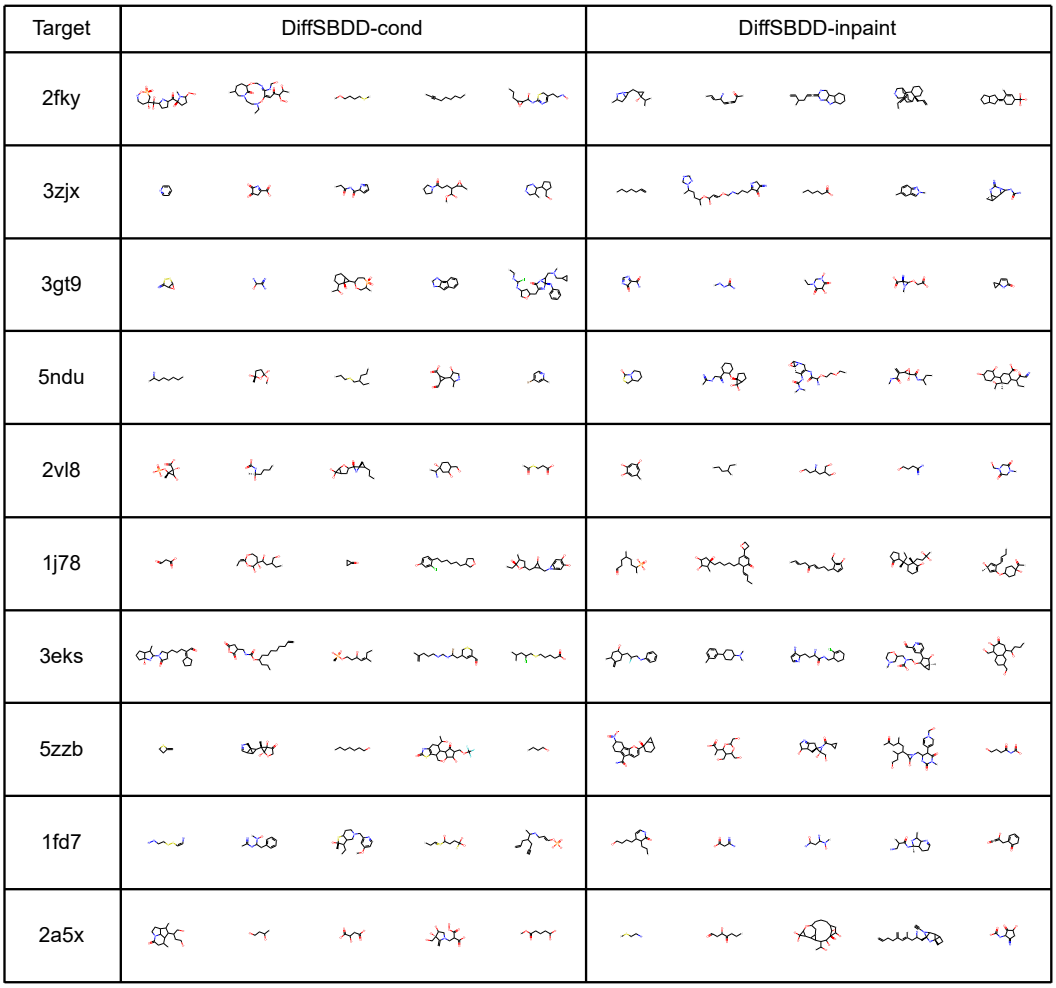

Figure 12: Generated molecules for 10 randomly chosen targets in the Binding MOAD test set. For each target-model pair, 5 randomly selected generated molecules are shown. C_α level proteins were used for both models.

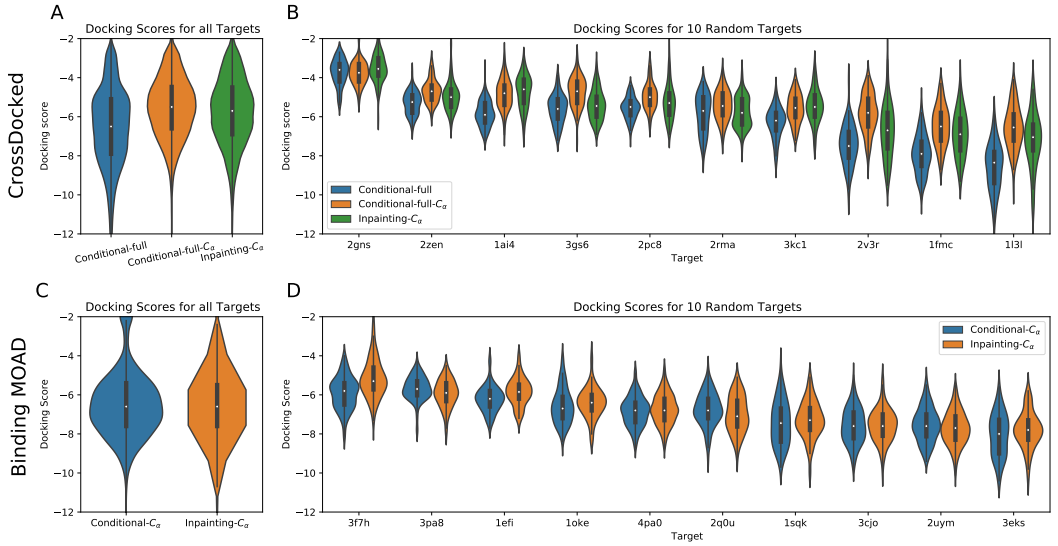

Figure 13: Docking scores of generated molecules for various methods trained on the CrossDocked (A-B) and Binding MOAD (C-D) datasets. (A) Violin plot of docking scores for all 3 methods trained using CrossDocked. (B) Same as before but for 10 randomly chosen targets sorted by mean score. (C) Violin plot of docking scores for all 2 methods trained using Binding MOAD. (D) Same as before but for 10 randomly chosen targets sorted by mean score.

cloud generation (Kingma et al., 2021; Kong et al., 2021; Luo & Hu, 2021). Recently, efforts have been made to utilize diffusion models for molecule design (Du et al., 2022b). Specifically, Hoogeboom et al. (2022) propose a diffusion model with an equivariant network that operates both on continuous atomic coordinates and categorical atom types to generate new molecules in 3D space. Torsional Diffusion (Jing et al., 2022) focuses on a conditional setting where molecular conformations (atomic coordinates) are generated from molecular graphs (atom types and bonds). Similarly, 3D diffusion models have been applied to generative design of larger biomolecular structures, such as antibodies (Luo et al., 2022) and other proteins (Anand & Achim, 2022; Trippe et al., 2022).

Structure-based Drug Design Structure-based Drug Design (SBDD) (Ferreira et al., 2015; Anderson, 2003) relies on the knowledge of the 3D structure of the biological target obtained either through experimental methods or high-confidence predictions using homology modelling (Kelley et al., 2015). Candidate molecules are then designed to bind with high affinity and specificity to the target using interactive software (Kalyaanamoorthy & Chen, 2011) and often human-based intuition (Ferreira et al., 2015). Recent advances in deep generative models have brought a new wave of research that model the conditional distribution of ligands given biological targets and thus enable *de novo* structure-based drug design. Most of recent work consider this task as a sequential generation problem and design a variety of generative methods including autoregressive models, reinforcement learning, etc., to generate ligands inside protein pockets atom by atom (Drotár et al., 2021; Luo et al., 2021; Li et al., 2021; Peng et al., 2022).

Geometric Deep Learning for Drug Discovery Geometric deep learning refers to incorporating geometric priors in neural architecture design that respects symmetry and invariance, thus reduces sample complexity and eliminates the need for data augmentation (Bronstein et al., 2021). It has been prevailing in a variety of drug discovery tasks from virtual screening to de novo drug design as symmetry widely exists in the representation of drugs. One line of work introduces graph and geometry priors and designs message passing neural networks and equivariant neural networks that are permutation- and translation-, rotation-, reflection-equivariant, respectively (Duvenaud et al., 2015; Gilmer et al., 2017; Satorras et al., 2021; Lapchevskyi et al., 2020; Du et al., 2022a), and has been widely used in representing biomolecules from small molecules to proteins (Atz et al., 2021) and solving downstream tasks such as molecular property prediction (Schütt et al., 2018; Klicpera et al., 2020), binding pose prediction (Stärk et al., 2022) or molecular dynamics (Batzner et al., 2022;

Holdijk et al., 2022). Another line of work focuses on generative design of new molecules (Du et al., 2022b;c). Specifically, they formulate molecule design as a graph or geometry generation problem and there are two strategies: one-shot generation that generates graphs (atom and bond features) in one step and sequential generation that generates them in a sequence of steps.