Journal Title Here, 2022, 1–14
doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

FlowDock: Geometric Flow Matching for Generative Protein-Ligand Docking and Affinity Prediction

Alex Morehead[1,*] **and Jianlin Cheng**[1]

[1]Department of Electrical Engineering & Computer Science, NextGen Precision Health, University of Missouri-Columbia, W1024 Lafferre Hall, 65211, Missouri, USA

[*]Corresponding author. acmwhb@missouri.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation

Powerful generative AI models of protein-ligand structure have recently been proposed, but few of these methods support both flexible protein-ligand docking and affinity estimation. Of those that do, none can directly model multiple binding ligands concurrently or have been rigorously benchmarked on pharmacologically relevant drug targets, hindering their widespread adoption in drug discovery efforts.

Results

In this work, we propose FlowDock, the first deep geometric generative model based on conditional flow matching that learns to directly map unbound (apo) structures to their bound (holo) counterparts for an arbitrary number of binding ligands. Furthermore, FlowDock provides predicted structural confidence scores and binding affinity values with each of its generated protein-ligand complex structures, enabling fast virtual screening of new (multi-ligand) drug targets. For the well-known PoseBusters Benchmark dataset, FlowDock outperforms single-sequence AlphaFold 3 with a 51% blind docking success rate using unbound (apo) protein input structures and without any information derived from multiple sequence alignments, and for the challenging new DockGen-E dataset, FlowDock outperforms single-sequence AlphaFold 3 and matches single-sequence Chai-1 for binding pocket generalization. Additionally, in the ligand category of the 16th community-wide Critical Assessment of Techniques for Structure Prediction (CASP16), FlowDock ranked among the top-5 methods for pharmacological binding affinity estimation across 140 protein-ligand complexes, demonstrating the efficacy of its learned representations in virtual screening.

Availability

Source code, data, and pre-trained models are available at `https://github.com/BioinfoMachineLearning/FlowDock`.

Key words: Generative AI model, flow matching, protein-ligand structure, binding affinity

1. Introduction

Interactions between proteins and small molecules (ligands) drive many of life's fundamental processes and, as such, are of great interest to biochemists, biologists, and drug discoverers. Historically, elucidating the structure, and therefore the function, of such interactions has required that considerable intellectual and financial resources be dedicated to determining the interactions of a single biomolecular complex. For example, techniques such as X-ray diffraction and cryo-electron microscopy have traditionally been effective in biomolecular structure determination, however, resolving even a single biomolecule's crystal structure can be extremely time and resource-intensive. Recently, new machine learning (ML) methods such as AlphaFold 3 (AF3) (Abramson et al., 2024) have been proposed for directly predicting the structure of an arbitrary biomolecule from its primary sequence, offering the potential to expand our understanding of life's molecules and their implications in disease, energy research, and beyond.

Although powerful models of general biomolecular structure are compelling, they currently do not provide one with an estimate of the binding affinity of a predicted protein-ligand complex, which may indicate whether a pair of molecules truly bind to each other *in vivo*. It is desirable to predict both the structure of a protein-ligand complex and the binding affinity between them via one single ML system (Dhakal et al., 2022). Moreover,

recent generative models of biomolecular structure are primarily based on noise schedules following Gaussian diffusion model methodology which, albeit a powerful modeling framework, lacks interpretability in the context of biological studies of molecular interactions. In this work, we aim to address these concerns with a new *state-of-the-art* hybrid (structure & affinity prediction) generative model called FlowDock for flow matching-based protein-ligand structure prediction and binding affinity estimation, which allows one to interpretably inspect the model's structure prediction trajectories to interrogate its common molecular interactions and to screen drug candidates quickly using its predicted binding affinities.

2. Related work

Molecular docking with deep learning. Over the last few years, deep learning (DL) algorithms (in particular geometric variants) have emerged as a popular methodology for performing end-to-end differentiable molecular docking. Models such as EquiBind (Stärk et al., 2022) and TankBind (Lu et al., 2022) initiated a wave of interest in researching graph-based approaches to modeling protein-ligand interactions, leading to many follow-up works. Important to note is that most of such DL-based docking models were designed to supplement conventional modeling methods for protein-ligand docking such as AutoDock Vina (Eberhardt et al., 2021) which are traditionally slow and computationally expensive to run for many protein-ligand complexes yet can achieve high accuracy with crystal input structures and ground-truth binding pocket annotations.

Generative biomolecular modeling. The potential of generative modeling in capturing intricate molecular details in structural biology such as protein-ligand interactions during molecular docking (Corso et al., 2022) has recently become a research focus of ambitious biomolecular modeling efforts such as AF3 (Abramson et al., 2024), with several open-source spin-offs of this algorithm emerging (Discovery et al., 2024; Wohlwend et al., 2024).

Flow matching. In the machine learning community, generative modeling with flow matching (Chen and Lipman, 2024; Tong et al., 2024) has recently become an appealing generalization of diffusion generative models (Ho et al., 2020; Karras et al., 2022), enabling one to transport samples between arbitrary distributions for compelling applications in computer vision (Esser et al., 2024), computational biology (Klein et al., 2024), and beyond. As a closely related concurrent work (as our method was developed for the CASP16 competition starting in May 2024 (CASP16-Organizers, 2024)), Corso et al. (2024b) recently introduced and evaluated an unbalanced flow matching procedure for pocket-based flexible docking. However, the authors' proposed approach mixes diffusion and flow matching noise schedules with geometric product spaces in an unintuitive manner, and neither source code nor data for this work are publicly available for benchmarking comparisons. In Section 3.3, we describe flow matching in detail.

Contributions. In light of such prior works, our contributions in this manuscript are as follows:

- We introduce the *first* simple yet state-of-the-art *hybrid* generative flow matching model capable of quickly and accurately predicting protein-ligand complex structures *and* their binding affinities, with source code and model weights freely available.
- We rigorously validate our proposed methodology using standardized benchmarking data for protein-ligand complexes, with our method ranking as a more accurate and generalizable structure predictor than (single-sequence) AF3.
- Our method ranked as a top-5 binding affinity predictor for the 140 pharmaceutically relevant drug targets available in the 2024 community-wide CASP16 ligand prediction competition.
- We release one of the largest ML-ready datasets of apo-to-holo protein structure mappings based on high-accuracy predicted protein structures, which enables training new models on comprehensive biological data for distributional biomolecular structure modeling.

3. Methods and materials

The goal of this work is to jointly predict protein-ligand complex structures and their binding affinities with minimal computational overhead to facilitate drug discovery. In Sections 3.1 and 3.2, we briefly outline how FlowDock achieves this and how its key notation is defined. We then describe FlowDock's training and sampling procedures in Sections 3.3-3.6.

3.1. Overview

Figure 1 illustrates how FlowDock uses geometric flow matching to predict flexible protein-ligand structures and binding affinities. At a high level, FlowDock accepts both (multi-chain) protein sequences and (multi-fragment) ligand SMILES strings as its primary inputs, which it uses to predict an unbound (apo) state of the protein sequences using ESMFold (Lin et al., 2023) and to sample from a harmonic ligand prior distribution (Jing et al., 2024) to initialize the ligand structures using biophysical constraints based on their specified bond graphs. Notably, users can also specify the initial protein structure using one produced by another bespoke method (e.g., AF3 which we use in certain experiments). With these initial structures representing the complex's state at time $t = 0$, FlowDock employs conditional flow matching to produce fast structure generation trajectories. After running a small number of integration timesteps (e.g., 40 in our experiments), the complex's state arrives at time $t = 1$, i.e., the model's estimate of the bound (holo) protein-ligand heavy-atom structure. At this point, FlowDock runs confidence and binding affinity heads to predict structural confidence scores (i.e., plDDT) and binding affinities of the predicted complex structure, to rank-order the model's generated samples.

3.2. Notation

Let $\boldsymbol{x}_0$ denote the unbound (apo) state of a protein-ligand complex structure, representing the heavy atoms of the protein and ligand structures as $\boldsymbol{x}_0^P \in \mathbb{R}^{N^P \times 3}$ and $\boldsymbol{x}_0^L \in \mathbb{R}^{N^L \times 3}$, respectively, where N^P and N^L are the numbers of protein and ligand heavy atoms. Similarly, we denote the corresponding bound (holo) state of the complex as $\boldsymbol{x}_1$. Further, let $\boldsymbol{s}^P \in \{1, \ldots, 20\}^{S^P}$ denote the type of each amino acid residue in the protein structure, where S^P represents the protein's sequence length. To generate bound (holo) structures, we define a flow model v_θ that integrates the ordinary differential equation (ODE) it defines from time $t = 0$ to $t = 1$.

3.3. Riemannian manifolds and conditional flow matching

In manifold theory, an n-dimensional manifold $\mathcal{M}$ represents a topological space equivalent to $\mathbb{R}^n$. In the context of Riemannian

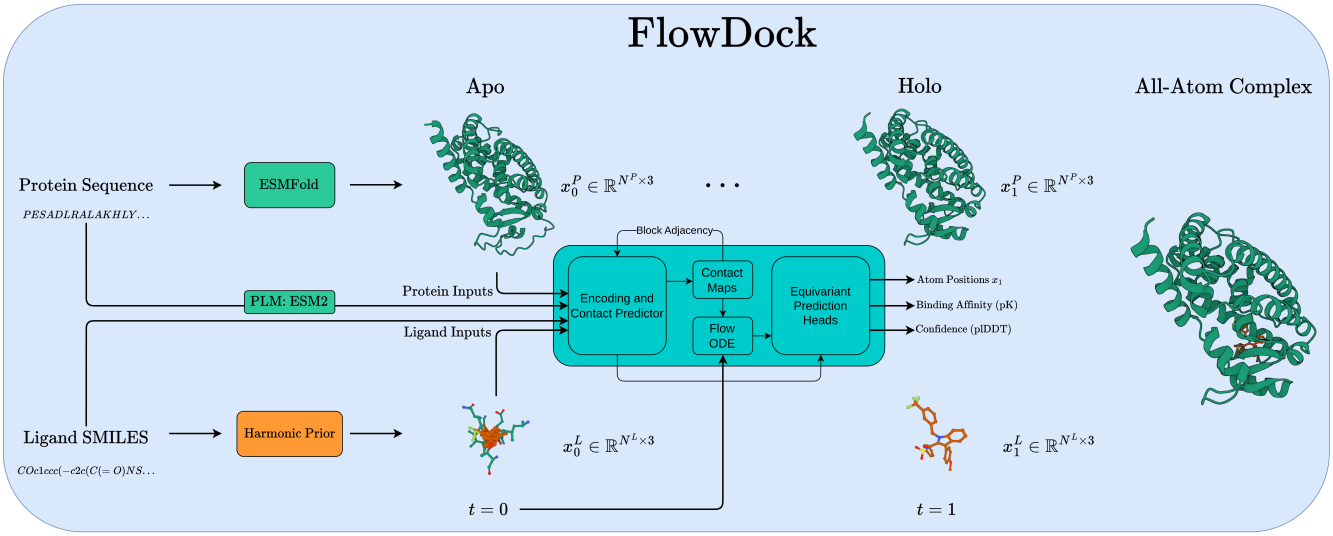


Fig. 1: An overview of biomolecular distribution modeling with FLOWDOCK.

manifold theory, each point $\boldsymbol{x} \in \mathcal{M}$ on a Riemannian manifold is associated with a tangent space $\mathcal{T}_{\boldsymbol{x}}$. Conveniently, a Riemannian manifold is equipped with a metric $g_{\boldsymbol{x}} : \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \times \mathcal{T}_{\boldsymbol{x}}\mathcal{M} \to \mathbb{R}$ that permits the definition of geometric quantities on the manifold such as distances and geodesics (i.e., shortest paths between two points on the manifold). Subsequently, Riemannian manifolds allow one to define on them probability densities $\int_{\mathcal{M}} \rho(\boldsymbol{x})d\boldsymbol{x} = 1$ where $\rho : \mathcal{M} \to \mathbb{R}_+$ are continuous, non-negative functions. Such probability densities give rise to interpolative probability paths $\rho_t : [0, 1] \to \mathbb{P}(\mathcal{M})$ between probability distributions $\rho_0, \rho_1 \in \mathbb{P}(\mathcal{M})$, where $\mathbb{P}(\mathcal{M})$ is defined as the space of probability distributions on $\mathcal{M}$ and the interpolation in probability space between distributions is indexed by the continuous parameter t.

Here, we refer to $\psi_t : \mathcal{M} \to \mathcal{M}$ as a *flow* on $\mathcal{M}$. Such a flow serves as a solution to the ODE: $\frac{d}{dt}\psi_t(\boldsymbol{x}) = u_t(\psi_t(\boldsymbol{x}))$ (Mathieu and Nickel, 2020) which allows one to *push forward* the probability trajectory $\rho_0 \to \rho_1$ to ρ_t using ψ_t as $\rho_t = [\psi_t]_{\#}(\rho_0)$, with $\psi_0(\boldsymbol{x}) = \boldsymbol{x}$ for $u : [0, 1] \times \mathcal{M} \to \mathcal{M}$ (i.e., a smooth time-dependent vector field (Bose et al., 2024)). This insight allows one to perform *flow matching* (FM) (Chen and Lipman, 2024) between ρ_0 and ρ_1 by learning a continuous normalizing flow (Papamakarios et al., 2021) to approximate the vector field u_t with the parametric v_θ. With $\rho_0 = \rho_{prior}$ and $\rho_1 = \rho_{data}$, we have that ρ_t advantageously permits *simulation-free* training. Although it is not possible to derive a closed form for u_t (which generates ρ_t) with the traditional flow matching (FM) training objective, a *conditional* flow matching (CFM) training objective remains tractable by marginalizing conditional vector fields as $u_t(\boldsymbol{x}) := \int_{\mathcal{M}} u_t(\boldsymbol{x}|\boldsymbol{z}) \frac{\rho_t(\boldsymbol{x}_t|\boldsymbol{z})q(\boldsymbol{z})}{\rho_t(\boldsymbol{x})} d\boldsymbol{z}$, where $q(\boldsymbol{z})$ represents one's chosen coupling distribution (by default the independent coupling $q(\boldsymbol{z}) = q(\boldsymbol{x}_0)q(\boldsymbol{x}_1)$) between $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ via the conditioning variable $\boldsymbol{z}$. For Riemannian CFM (RCFM) (Chen and Lipman, 2024), the corresponding training objective, with $t \sim \mathcal{U}(0, 1)$, is:

$$\mathcal{L}_{RCFM}(\theta) = \mathbb{E}_{t,q(\boldsymbol{z}),\rho_t(\boldsymbol{x}_t|\boldsymbol{z})} \|v_\theta(\boldsymbol{x}_t, t) - u_t(\boldsymbol{x}_t|\boldsymbol{z})\|_g^2, \quad (1)$$

where Tong et al. (2024) have fortuitously shown that the gradients of FM and CFM are identical. As such, to transport samples of the prior distribution ρ_0 to the target (data) distribution ρ_1, one can sample from ρ_0 and use v_θ to run the corresponding ODE forward in time. In the remainder of this work, we will focus specifically on the 3-manifold $\mathbb{R}^3$.

3.4. Prior distributions

With flow matching defined, in this section, we describe how we use a bespoke mixture of prior distributions (ρ_0^P and ρ_0^L) to sample initial (unbound) protein and ligand structures for binding (holo) structure generation targeting our data distribution of crystal protein-ligand complex structures ρ_1. In Section 4.1, we ablate this mixture to understand its empirical strengths.

ESMFold protein prior. To our best knowledge, FLOWDOCK is among the *first* methods-concurrently with Corso et al. (2024b)-to explore using structure prediction models with flow matching to represent the unbound state of an arbitrary protein sequence. In contrast to Corso et al. (2024b), we formally define a *distribution* of unbound (apo) protein structures using the single-sequence ESMFold model as $\rho_0^P(\boldsymbol{x}_0^P) \propto \text{ESMFold}(\boldsymbol{s}^P) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma)$, which encourages our model to learn more than a strict mapping between protein apo and holo point masses. Based on previous works developing protein generative models (Dauparas et al., 2022), during training we apply $\epsilon \sim \mathcal{N}(0, \sigma = 1e^{-4})$ to both $\boldsymbol{x}_0^P$ and $\boldsymbol{x}_1^P$ to discourage our model from overfitting to computational or experimental noise in its training data. It is important to note that this additive noise for protein structures is not a general substitute for generating a full conformational ensemble of each protein, but to avoid the excessively high computational resource requirements of running protein dynamics methods such as AlphaFlow (Jing et al., 2024) for each protein, we empirically find noised ESMFold structures to be a suitable surrogate.

Harmonic ligand prior. Inspired by the FlowSite model for multi-ligand binding site design (Stark et al., 2024), FLOWDOCK samples initial ligand conformations using a harmonic prior distribution constrained by the bond graph defined by one's specified ligand SMILES strings. This prior can be sampled as a modified Gaussian distribution via $\rho_0^L(\boldsymbol{x}_0^L) \propto exp(-\frac{1}{2}\boldsymbol{x}_0^{L^T}\boldsymbol{L}\boldsymbol{x}_0^L)$

where $\boldsymbol{L}$ denotes a ligand bond graph's Laplacian matrix defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$, with $\boldsymbol{A}$ being the graph's adjacency matrix and $\boldsymbol{D}$ being its degree matrix. Similarly to our ESMFold protein prior, we subsequently apply $\epsilon \sim \mathcal{N}(0, \sigma = 1e^{-4})$ to $\boldsymbol{x}_1^L$ during training.

3.5. Training

We describe FLOWDOCK's structure parametrization, optimization procedure, and the curation and composition of its new training dataset in the following sections. Further, we provide training and inference pseudocode in Appendix A of our Supplementary Materials.

Parametrizing protein-ligand complexes with geometric flows. Based on our experimental observations of the difficulty of scaling up intrinsic generative models (Corso, 2023) that operate on geometric product spaces, FLOWDOCK instead parametrizes 3D protein-ligand complex structures as attributed geometric graphs (Joshi et al., 2023) representing the heavy atoms of each complex's protein and ligand structures. The main benefit of a heavy atom parametrization is that it can considerably simplify the optimization of a flow model v_θ by allowing one to define its primary loss function as simply as a CondOT path (Pooladian et al., 2023; Jing et al., 2024):

$$\mathcal{L}_{\mathbb{R}^3}(\theta) = \mathbb{E}_{t, q(\boldsymbol{z}), \rho_t(\boldsymbol{x}_t|\boldsymbol{z})} \| v_\theta(\boldsymbol{x}_t, t) - \boldsymbol{x}_1) \|^2, \tag{2}$$

with the conditional probability path ρ_t chosen as

$$\rho_t(\boldsymbol{x}|\boldsymbol{z}) = \rho_t(\boldsymbol{x}|\boldsymbol{x}_0, \boldsymbol{x}_1) = (1-t) \cdot \boldsymbol{x}_0 + t \cdot \boldsymbol{x}_1, \quad \boldsymbol{x}_0 \sim \rho_0(\boldsymbol{x}_0) \tag{3}$$

The challenge introduced by this atomic parametrization is that it necessitates the development of an efficient neural architecture that can scalably process all-atom input structures without the exhaustive computational overhead of generative models such as AF3. Fortunately, one such architecture satisfies this requirement, namely, one recently introduced by Qiao et al. (2024) with the NeuralPLexer model which encodes protein language model (PLM) sequence embeddings and ligand SMILES strings to iteratively decode block diagonal contact maps to condition a flow ODE for equivariant coordinates and auxiliary predictions. As such, inspired by how the AlphaFlow model was fine-tuned from the base AlphaFold 2 (AF2) architecture using flow matching, to train FLOWDOCK we explored fine-tuning the NeuralPLexer architecture to represent our vector field estimate v_θ as illustrated in Figure 1. Uniquely, we empirically found this idea to work best by fine-tuning the architecture's score head, which was originally trained with a denoising score matching objective for *diffusion*-based structure sampling, instead using Eqs. 2 and 3. Moreover, we fine-tune all of NeuralPLexer's remaining intermediate weights and prediction heads including a dedicated confidence head redesigned to predict binding affinities, with the exception of its original confidence head which remains frozen at all points during training.

PDBBind-E Data Curation. To train FLOWDOCK with resolved protein-ligand structures and binding affinities, we prepared PDBBind-E, an enhanced version of the PDBBind 2020-based training dataset proposed by Corso et al. (2024a) for training recent DL docking methods such as DiffDock-L. To curate PDBBind-E, we collected 17,743 crystal complex structures contained in the PDBBind 2020 dataset and 47,183 structures of the Binding MOAD (Hu et al., 2005) dataset splits introduced by Corso et al. (2024a) (n.b., which maintain the validity of our benchmarking results in Section 4 according to time and ligand-based similarity cutoffs) and predicted the structure of these (multi-chain) protein sequences in each dataset split using ESMFold. To optimally align each predicted protein structure with its corresponding crystal structure, we performed a weighted structural alignment optimizing for the distances of the predicted protein residues' Cα atoms to the crystal heavy atom positions of the complex's binding ligand, similar to (Corso et al., 2024a). After dropping complexes for which the crystal structure contained protein sequence gaps caused by unresolved residues, the total number of PDBBind and Binding MOAD predicted complex structures remaining was 17,743 and 46,567, respectively.

Generalized unbalanced flow matching. We empirically observed the challenges of naively training flexible docking models like FLOWDOCK without any adjustments to the sampling of their training data. Accordingly, we concurrently developed a generalized version of *unbalanced* flow matching (Corso et al., 2024b) by defining our coupling distribution $q(\boldsymbol{z})$ as

$$q(\boldsymbol{x}_0, \boldsymbol{x}_1) \propto q_0(\boldsymbol{x}_0) q_1(\boldsymbol{x}_1) \mathbb{I}_{c(\boldsymbol{x}_0, \boldsymbol{x}_1) \in c_{\mathbb{A}}}, \tag{4}$$

where $c_{\mathbb{A}}$ is defined as a set of apo-to-holo assessment filters measuring the structural similarity of the unbound (apo) and bound (holo) protein structures (n.b., not simply their binding pockets) in terms of their root mean square deviation (RMSD) and TM-score (Zhang and Skolnick, 2004) following optimal structural alignment (as used in constructing PDBBind-E). Effectively, we sample independent examples from q_0 and q_1 and reject these paired examples if $c(\boldsymbol{x}_0, \boldsymbol{x}_1) < c_{\mathbb{A}_{TM}}$ or $c(\boldsymbol{x}_0, \boldsymbol{x}_1) \geq c_{\mathbb{A}_{RMSD}}$ (n.b., we use $c_{\mathbb{A}_{TM}} = 0.7$ and $c_{\mathbb{A}_{RMSD}} = 5$Å as well as other length-based criteria in our experiments, please see our code for full details).

3.6. Sampling

By default, we apply $i = 40$ timesteps of an Euler solver to integrate FLOWDOCK's learned ODE v_θ forward in time for binding (holo) structure generation. Specifically, to generate structures, we propose to integrate a Variance Diminishing ODE (VD-ODE) that uses v_θ as

$$\boldsymbol{x}_{n+1} = clamp(\frac{1-s}{1-t} \cdot \eta) \cdot \boldsymbol{x}_n + clamp((1 - \frac{1-s}{1-t}) \cdot \eta) \cdot v_\theta(\boldsymbol{x}_n, t), \tag{5}$$

where n represents the current integer timestep, allowing us to define $t = \frac{n}{i}$ and $s = \frac{n+1}{i}$; $\eta = 1.0$ in our experiments; and *clamp* ensures both the LHS and RHS of Eq. 5 are lower and upper bounded by $1e^{-6}$ and $1 - 1e^{-6}$, respectively. We experimented with different values of η yet ultimately settled on 1.0 since this yielded FLOWDOCK's best performance for structure and affinity prediction. Intuitively, this VD-ODE solver limits the high levels of variance present in the model's predictions v_θ during early timesteps by sharply interpolating towards v_θ in later timesteps.

4. Results

4.1. PoseBench protein-ligand docking

PoseBusters Benchmark set. In Figures 2 and 3, we illustrate the performance of each baseline method for protein-ligand docking and protein conformational modification with the commonly-used PoseBusters Benchmark set (Buttenschoen et al., 2024), provided by version 0.6.0 of the PoseBench protein-ligand benchmarking suite (Morehead et al., 2024), which consists of 308 distinct protein-ligand complexes released after 2020. It is

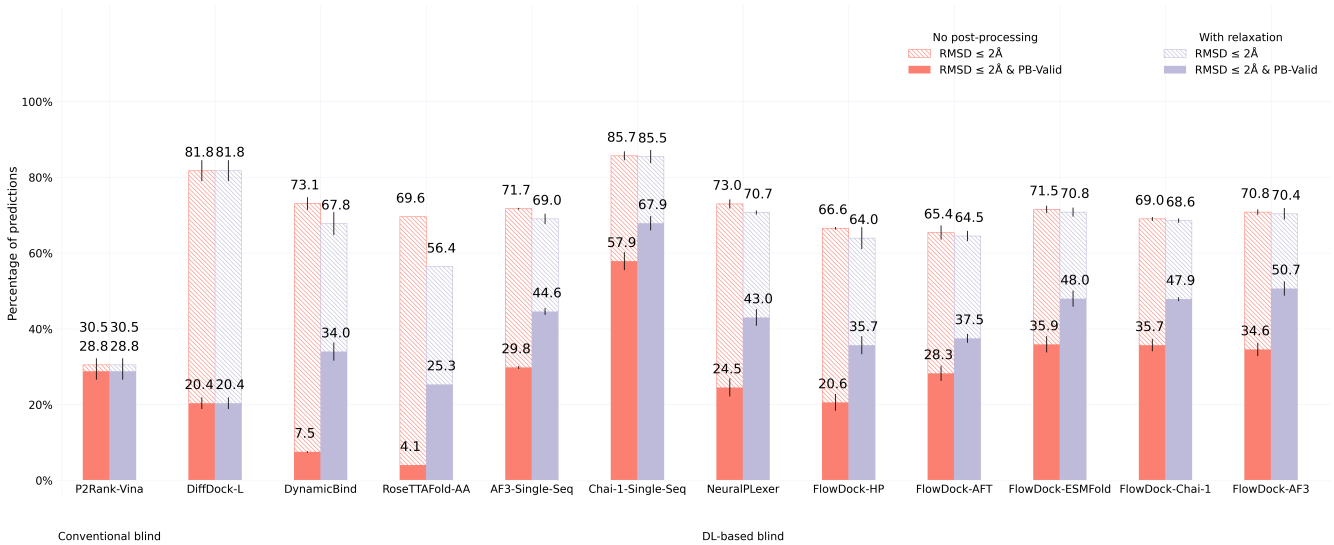
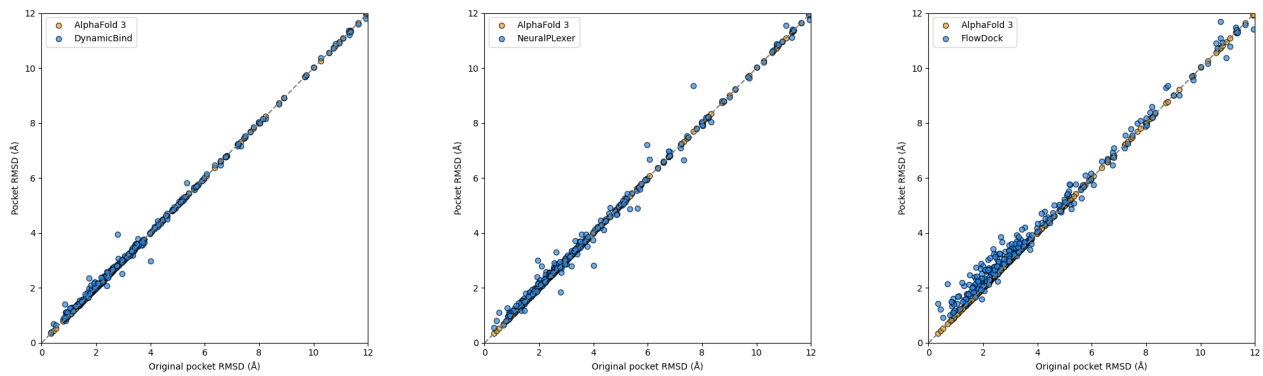


Fig. 2: Protein-ligand docking success rates of each baseline method on the PoseBusters Benchmark set (n=308). Error bars: 3 runs.



DynamicBind
(Binding Pocket Conformation RMSDs)

NeuralPLexer
(Binding Pocket Conformation RMSDs)

FlowDock
(Binding Pocket Conformation RMSDs)

Fig. 3: Comparison of each flexible docking method's protein conformational changes made for the PoseBusters Benchmark set (n=308).

important to note that this benchmarking set can be considered a moderately difficult challenge for methods trained on recent collections of data derived from the Protein Data Bank (PDB) (Bank, 1971) such as PDBBind 2020 (Liu et al., 2015), as all of these 308 protein-ligand complexes are not contained in the most common training splits of such PDB-based data collections (Buttenschoen et al., 2024) (with the exception of AF3 which uses a cutoff date of September 30, 2021). Moreover, as described by Buttenschoen et al. (2024), a subset of these complexes also have very low protein sequence similarity to such training splits.

Figure 2 shows that FLOWDOCK consistently improves over the original NeuralPLexer model's docking success rate in terms of its structural and chemical accuracy (as measured by the RMSD $\leq$ 2Å & PB-Valid metric (Buttenschoen et al., 2024)) and inter-run stability (as measured by the error bars listed). Notably, FLOWDOCK achieves a 10% higher docking success rate than NeuralPLexer without any structural energy minimization driven by molecular dynamics software (Eastman et al., 2017), and with energy minimization its docking success rate increases to 51%, outperforming single-sequence AF3 and achieving second-best performance on this dataset compared to single-sequence Chai-1 (Discovery et al., 2024). Important to note is that Chai-1, like AF3, is a 10x larger model trained for one month using 128 NVIDIA A100 80GB GPUs on more than twice as much data in the PDB deposited up to 2021, whereas FLOWDOCK is trained using only 4 80GB H100 GPUs for one week, representing a 32x reduction in GPU hours required for training. Additionally, FLOWDOCK outperforms the *hybrid* flexible docking method DynamicBind (Lu et al., 2024) by more than 16%, which is a comparable model in terms of its size, training, and downstream capabilities for drug discovery. Our results with ablated versions of FLOWDOCK trained instead with a protein harmonic prior (FLOWDOCK-HP) or with affinity prediction frozen until a fine-tuning phase (FLOWDOCK-AFT) highlight that the protein ESMFold prior the base FLOWDOCK model employs has imbued it with meaningful structural representations for accurate

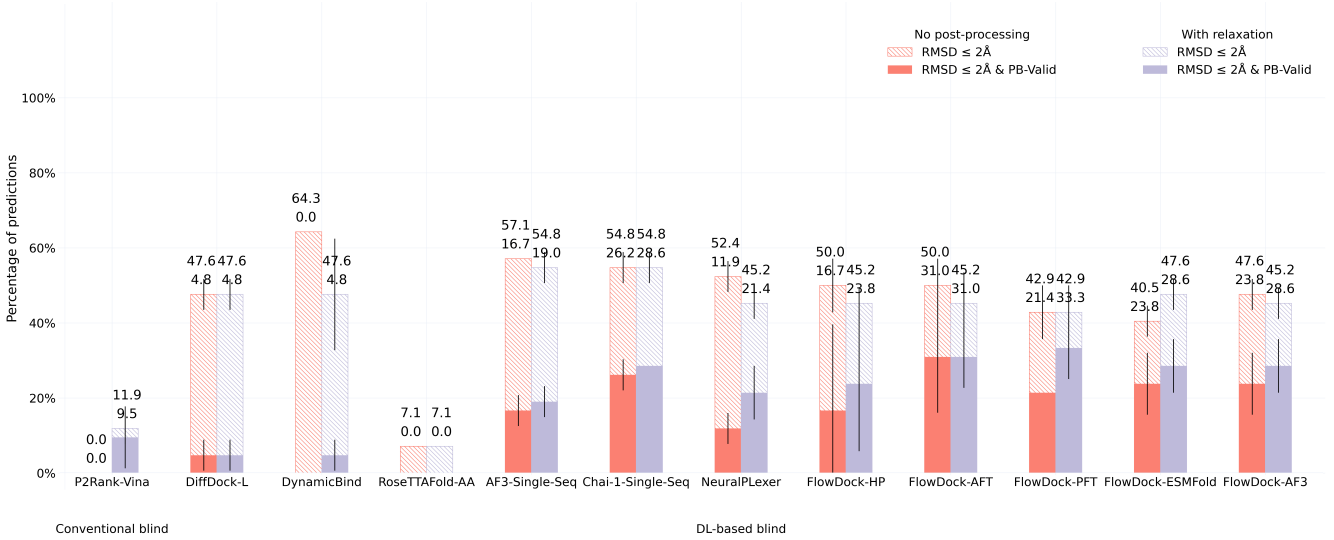
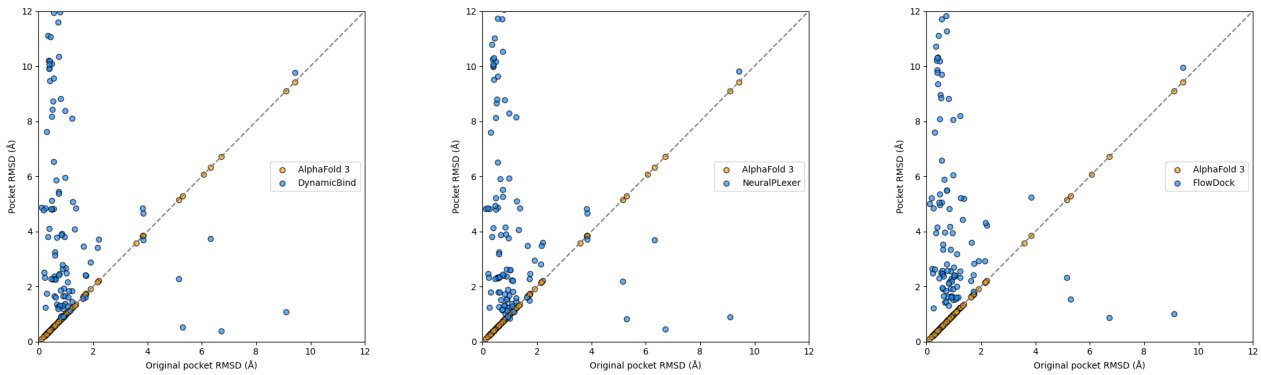


Fig. 4: Protein-ligand docking success rates of each baseline method on the DockGen-E set (n=14). Error bars: 3 runs.



DynamicBind
(Binding Pocket Conformation RMSDs)

NeuralPLexer
(Binding Pocket Conformation RMSDs)

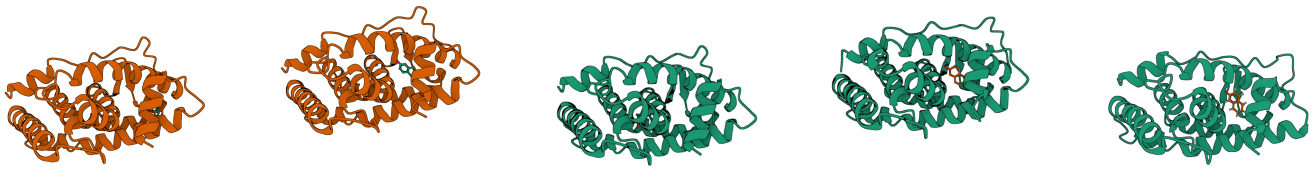
FlowDock
(Binding Pocket Conformation RMSDs)

Fig. 5: Comparison of each flexible docking method's protein conformational changes made for the DockGen-E set (n=122).

ligand binding structure prediction that are robust to changes in the source method of FlowDock's predicted protein input structures (e.g., FlowDock-ESMFold vs. FlowDock-Chai-1 vs. FlowDock-AF3), providing users with multiple structure prediction options (e.g., ESMFold for faster and commercially available prediction inputs).

A surprising finding illustrated in Figure 3 is that no method can consistently improve the binding pocket RMSD of AF3's initial protein structural conformations, which contrasts with the results originally reported for flexible docking methods such as DynamicBind which used structures predicted by AF2 (Jumper et al., 2021) in its experiments. From this figure, we observe that DynamicBind and NeuralPLexer both infrequently modify AF3's initial binding pocket structure, whereas FlowDock often modifies the pocket structure during ligand binding. The former two methods occasionally improve largely-correct initial pocket conformations by ∼1Å, whereas FlowDock primarily does so for mostly-incorrect initial pockets.

DockGen-E set. To assess the generalization capabilities of each baseline method, in Figures 4 and 5, we report each method's protein-ligand docking and protein conformational modification performance for the novel (i.e., naturally rare) protein binding pockets found in the new DockGen-E dataset from PoseBench. Each of DockGen-E's protein-ligand complexes represents a distinct binding pocket that facilitates a unique biological function described by its associated ECOD domain identifier (Corso et al., 2024a). As our results for the DockGen-E dataset show in Figure 4, most DL-based docking or structure prediction methods have likely not been trained or overfitted to these binding pockets, as this dataset's best docking success rate achieved by any method is approximately 33%, much lower than the 68% best docking success rate achieved for the PoseBusters Benchmark set. We find further support for this phenomenon in Figure 5, where we see that all DL-based flexible docking methods find it challenging to avoid degrading the initial binding pocket state predicted by AF3 yet all methods can *restore* a handful of AF3 binding pockets to their



DynamicBind Init. (ESMFold Structure) | **DynamicBind Final** (Affinity / plDDT: 5.72 / 0.83) | **FlowDock Init.** (ESMFold Structure) | **FlowDock Final** (Affinity / plDDT: 6.05 / 0.78) | **6I67 Crystal** (True Affinity: 6.70)

Fig. 6: Comparison of DYNAMICBIND and FLOWDOCK's predicted structures (w/o hydrogens) and crystal PDBBind test example 6I67.

Table 1. Computational resource requirements. The average structure prediction runtime (in seconds) and peak memory usage (in GB) of baseline methods on a 25% subset of the Astex Diverse dataset (Hartshorn et al., 2007) using an NVIDIA 80GB A100 GPU for benchmarking (with baselines taken from (Morehead et al., 2024)). The symbol - denotes a result that could not be estimated.

Method	Runtime (s)	CPU Memory Usage (GB)	GPU Memory Usage (GB)
P2Rank-Vina	1,283.70	9.62	0.00
DiffDock-L	88.33	8.99	70.42
DynamicBind	146.99	5.26	18.91
RoseTTAFold-All-Atom	3,443.63	55.75	72.79
AF3	3,049.41	-	-
AF3-Single-Seq	58.72	-	-
Chai-1-Single-Seq	114.86	58.49	56.21
NeuralPLexer	29.10	11.19	31.00
FlowDock	39.34	11.98	25.61

Table 2. Binding affinity estimation using PDBBind test set. For all methods, binding affinities were predicted in *one shot* using the commonly-used 363 PDBBind (ligand and time-split) test complexes (with splits and baselines from Lu et al. (2024)). Results for FLOWDOCK are reported as the mean and standard error of measurement ($n = 3$) of each metric over three independent runs. Note that, for historical reasons, the results for each version of FLOWDOCK were obtained using ESMFold predicted protein input structures.

Method	Pearson (↑)	Spearman (↑)	RMSE (↓)	MAE (↓)
GIGN	0.286	0.318	1.736	1.330
TransformerCPI	0.470	0.480	1.643	1.317
MONN	0.545	0.535	1.371	1.103
TankBind	0.597	0.610	1.436	1.119
DynamicBind (One-Shot)	0.665	0.634	1.301	1.060
FLOWDOCK-HP	0.577 ± 0.001	0.560 ± 0.001	1.516 ± 0.001	1.196 ± 0.002
FLOWDOCK-AFT	0.663 ± 0.003	0.624 ± 0.003	1.392 ± 0.005	1.113 ± 0.003
FlowDock	$\mathbf{0.705 \pm 0.001}$	$\mathbf{0.674 \pm 0.002}$	1.363±0.003	1.067±0.003

bound (holo) form. This suggests that all DL methods (some more so than others) struggle to generalize to novel binding pockets, yet FLOWDOCK achieves top performance in this regard by tying with single-sequence Chai-1. Further, to address this generalization issue, our preliminary results fine-tuning FLOWDOCK for 48 hours using the new, diverse PLINDER (Durairaj et al., 2024) dataset (i.e., FLOWDOCK-PFT), where we use the dataset's crystal apo-to-holo mapped protein-ligand complex structures contained within its default PL50 training split and deposited in the PDB before 2018, suggest that comprehensively training new DL methods on diverse protein-ligand binding structures is a promising direction towards generalizable docking.

Computational resources. To formally measure the computational resources required to run each baseline method, in Table 1 we list the average runtime (in seconds) and peak CPU (GPU) memory usage (in GB) consumed by each method when running them on a 25% subset of the Astex Diverse dataset (Hartshorn et al., 2007) (baseline results taken from Morehead et al. (2024)). Here, we notably find that FLOWDOCK provides the second lowest computational runtime and GPU memory usage compared to all other DL methods, enabling one to use commodity computing hardware to quickly screen new drug candidates using combinations of FLOWDOCK's predicted heavy-atom structures, confidence scores, and binding affinities.

4.2. PDBBind binding affinity estimation

In this section, we explore binding affinity estimation with FLOWDOCK using the PDBBind 2020 test dataset (n=363) originally curated by (Stärk et al., 2022), with benchmarking results shown in Table 2. Popular affinity prediction baselines listed in Table 2 such as TankBind (Lu et al., 2022) and DynamicBind (Lu et al., 2024) demonstrate that accurate affinity estimations are possible using hybrid DL models of protein-ligand structures and affinities. Here, we find that, as a hybrid deep generative model, FLOWDOCK provides the best Pearson and Spearman's correlations compared to all other baselines including FLOWDOCK-HP (a fully harmonic variant of FLOWDOCK) and FLOWDOCK-AFT (an ESMFold prior variant trained first for structure prediction and then with affinity fine-tuning) and produces compelling root mean squared error (RMSE) and mean absolute error (MAE) rates compared to the previous state-of-the-art method DynamicBind. Referencing Table 1, we further note that FLOWDOCK's average computational runtime per protein-ligand complex is more than 3 times lower than that of DynamicBind, demonstrating that FLOWDOCK, to our best knowledge, is currently the *fastest* binding affinity estimation method to match or exceed DynamicBind's level of accuracy for predicting binding affinities using the PDBBind 2020 dataset.

In Figure 6, we provide an illustrative example of a protein-ligand complex in the PDBBind test set (6I67) for which FLOWDOCK predicts notably more accurate complex structural motions and binding affinity values than the hybrid DL method DynamicBind, importantly recognizing that the right-most protein loop domain should be moved further to the right to facilitate ligand binding (see Appendix B of our Supplementary Materials for an example of one of FLOWDOCK's interpretable structure generation trajectories). One should note that, for historical reasons, our experiments with this PDBBind-based test set employed protein structures predicted by ESMFold (not AF3). In the next section, we explore an even more practical application of FLOWDOCK's fast and accurate structure and binding affinity predictions in the CASP16 ligand prediction competition.

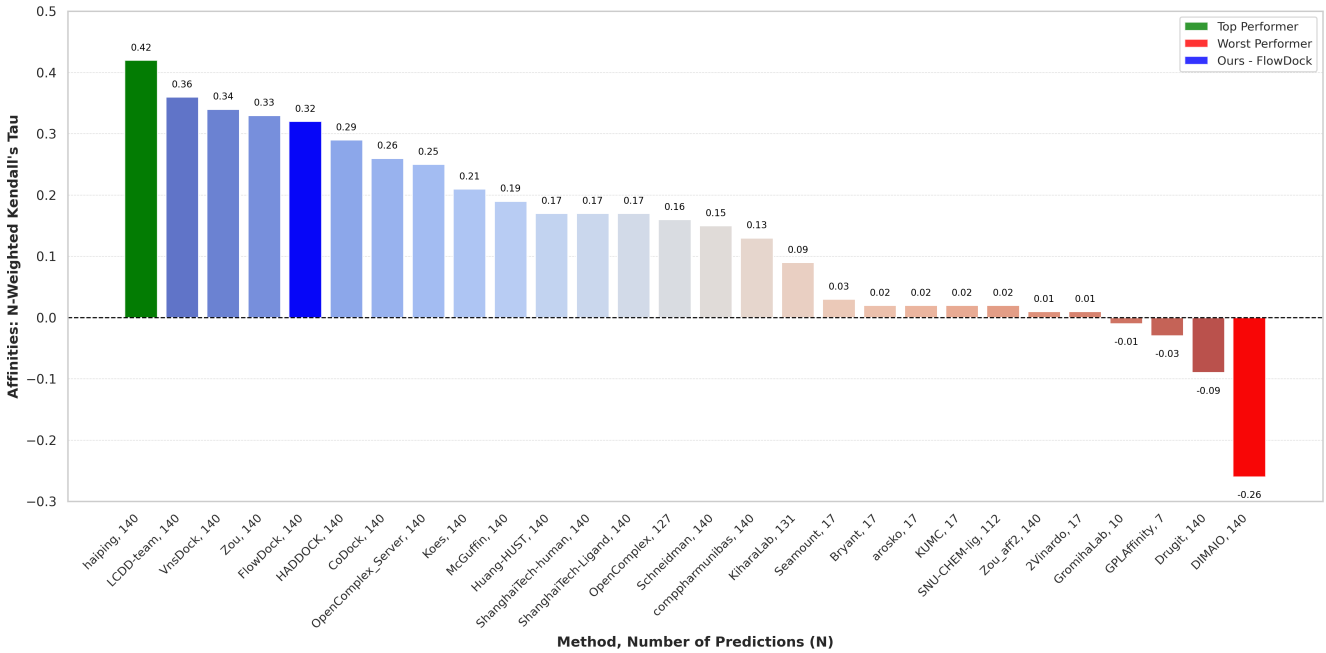


Fig. 7: Protein-ligand binding affinity prediction rankings for the CASP16 ligand prediction category (n=140).

4.3. CASP16 protein-ligand binding affinity prediction

In Figure 7, we illustrate the performance of each predictor group for blind protein-ligand binding affinity prediction in the ligand category of the CASP16 competition held in summer 2024, in which pharmaceutically relevant binding ligands were the primary focus of this competition. Notably, FlowDock is the *only* hybrid (structure & affinity prediction) ML method represented among the top-5 predictors, demonstrating the robustness of its knowledge of protein-ligand interactions. Namely, all other top prediction methods were trained specifically for binding affinity estimation assuming a predicted or crystal complex structure is provided. In contrast, in CASP16, we demonstrated the potential of using FlowDock to predict *both* protein-ligand structures and binding affinities and using its top-5 predicted structures' structural confidence scores to rank-order its top-5 binding affinity predictions (see Appendices C and D of our Supplementary Materials for FlowDock's e.g., CASP16 structure prediction results). Ranked 5th for binding affinity estimation, these results of the CASP16 competition demonstrate that this dual approach of predicting protein-ligand structures and binding affinities with a single DL model (FlowDock) yields compelling performance for virtual screening of pharmaceutically interesting molecular compounds.

5. Conclusion

In this work, we have presented FlowDock, a novel, state-of-the-art deep generative flow model for fast and accurate (hybrid) protein-ligand binding structure and affinity prediction. Benchmarking results suggest that FlowDock achieves structure prediction results better than single-sequence AF3 and comparable to single-sequence Chai-1 and outperforms existing hybrid models like DynamicBind across a range of binding ligands. Lastly, we have demonstrated the pharmaceutical virtual screening potential of FlowDock in the CASP16 ligand prediction competition, where it achieved top-5 performance. Future work could include retraining the model on larger and more diverse clusters of protein-ligand complexes, experimenting with new ODE solvers, or scaling up its parameter count to see if it displays any scaling law behavior for structure or affinity prediction. As a deep generative model for structural biology made available under an MIT license, we believe FlowDock takes a notable step forward towards fast, accurate, and broadly applicable modeling of protein-ligand interactions.

6. Conflict of interest

No conflicts of interest are declared.

7. Author contributions statement

A.M. and J.C. conceived the research. A.M. conducted the experiment(s). J.C. acquired funding to support this work. A.M. and J.C. analyzed the results and wrote the manuscript.

8. Funding

The authors thank the anonymous reviewers for their valuable suggestions. This work was supported by a U.S. NSF grant (DBI2308699) and a U.S. NIH grant (R01GM093123) awarded to J.C. Additionally, this work was performed using computing infrastructure provided by Research Support Services at the University of Missouri-Columbia (DOI: 10.32469/10355/97710).

References

J. Abramson, J. Adler, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, pages 1–3, 2024.

P. D. Bank. Protein data bank. *Nature New Biol*, 233(223): 10–1038, 1971.

J. Bose, T. Akhound-Sadegh, et al. Se(3)-stochastic flow matching for protein backbone generation. In *The Twelfth International Conference on Learning Representations*, 2024.

M. Buttenschoen, G. M. Morris, and C. M. Deane. Posebusters: Ai-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024.

CASP16-Organizers. Casp16 abstracts. *CASP16*, 2024. URL `https://predictioncenter.org/casp16/doc/CASP16_Abstracts.pdf#page=171.08`.

R. T. Chen and Y. Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024.

G. Corso. *Modeling molecular structures with intrinsic diffusion models*. Massachusetts Institute of Technology, 2023.

G. Corso, H. Stärk, et al. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

G. Corso, A. Deng, et al. Deep confident steps to new pockets: Strategies for docking generalization. In *The Twelfth International Conference on Learning Representations*, 2024a.

G. Corso, V. R. Somnath, et al. Flexible docking via unbalanced flow matching. In *ICML Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024b.

J. Dauparas, I. Anishchenko, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615): 49–56, 2022.

A. Dhakal, C. McKay, and J. Cheng. Artificial intelligence in the prediction of protein–ligand interactions: recent advances and future directions. *Briefings in Bioinformatics*, 23(1):bbab476, 2022.

C. Discovery, J. Boitreaud, et al. Chai-1: Decoding the molecular interactions of life. *bioRxiv*, pages 2024–10, 2024.

J. Durairaj, Y. Adeshina, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. In *ICML Workshop ML for Life and Material Science: From Theory to Industry Applications*, 2024.

P. Eastman, J. Swails, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):e1005659, 2017.

J. Eberhardt, D. Santos-Martins, et al. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61 (8):3891–3898, 2021.

P. Esser, S. Kulal, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

M. J. Hartshorn, M. L. Verdonk, et al. Diverse, high-quality test set for the validation of protein- ligand docking performance. *Journal of medicinal chemistry*, 50(4):726–741, 2007.

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

L. Hu, M. L. Benson, et al. Binding moad (mother of all databases). *Proteins: Structure, Function, and Bioinformatics*, 60(3):333–340, 2005.

B. Jing, B. Berger, and T. Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *Forty-first International Conference on Machine Learning*, 2024.

C. K. Joshi, C. Bodnar, et al. On the expressive power of geometric graph neural networks. In *International conference on machine learning*, pages 15330–15355. PMLR, 2023.

J. Jumper, R. Evans, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

T. Karras, M. Aittala, et al. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

L. Klein, A. Krämer, and F. Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Z. Lin, H. Akin, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379 (6637):1123–1130, 2023.

Z. Liu, Y. Li, et al. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

W. Lu, Q. Wu, et al. Tankbind: Trigonometry-aware neural networks for drug-protein binding structure prediction. *Advances in neural information processing systems*, 35:7236–7249, 2022.

W. Lu, J. Zhang, et al. Dynamicbind: Predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nature Communications*, 15(1):1071, 2024.

E. Mathieu and M. Nickel. Riemannian continuous normalizing flows. *Advances in Neural Information Processing Systems*, 33: 2503–2515, 2020.

A. Morehead, N. Giri, J. Liu, and J. Cheng. Deep learning for protein-ligand docking: Are we there yet? *ICML AI4Science Workshop*, 2024.

G. Papamakarios, E. Nalisnick, et al. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

A.-A. Pooladian, H. Ben-Hamu, et al. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, pages 28100–28127. PMLR, 2023.

Z. Qiao, W. Nie, et al. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6(2):195–208, 2024.

H. Stärk, O. Ganea, et al. Equibind: Geometric deep learning for drug binding structure prediction. In *International conference on machine learning*, pages 20503–20521. PMLR, 2022.

H. Stark, B. Jing, et al. Harmonic self-conditioned flow matching for joint multi-ligand docking and binding site design. In *Forty-first International Conference on Machine Learning*, 2024.

A. Tong, K. Fatras, and others. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Expert Certification.

J. Wohlwend, G. Corso, et al. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, pages 2024–11, 2024.

Y. Zhang and J. Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A. Morehead, J. Liu, P. Neupane, N. Giri, and J. Cheng. Protein-ligand structure and affinity prediction in casp16 using a geometric deep learning ensemble and flow matching. *Authorea*, 2025.

A. Geometric flow matching training and inference

We characterize FlowDock's training and sampling procedures in Sections 3.5 (Training) and 3.6 (Sampling) of the main text, respectively. To further illustrate how training and inference with FlowDock work, in Algorithms 1 and 2 we provide the corresponding pseudocode. For more details, please see our accompanying source code at `https://github.com/BioinfoMachineLearning/FlowDock`.

Algorithm 1 Training

Require: Training examples of binding site-aligned apo (holo) protein (ligand) structures, protein sequences, ligand SMILES strings, and binding affinities $\{(X^P_{a_i}, X^P_{h_i}, X^L_{h_i}, S_i, M_i, B_i)\}$

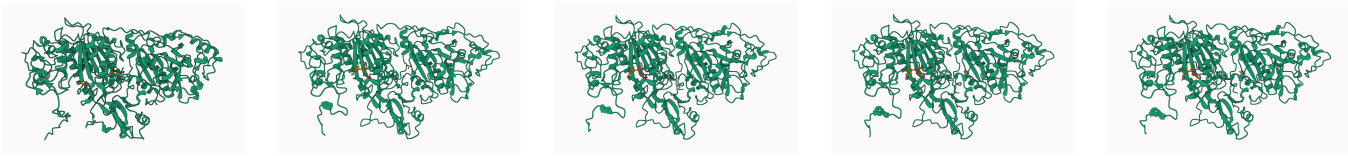
for all $(X^P_{a_i}, X^P_{h_i}, X^L_{h_i}, S_i, M_i, B_i)$ **do**
  Extract $x^P_1, x^L_1 \leftarrow \text{HeavyAtoms}(X^P_{h_i}, X^L_{h_i})$
  Sample $x^P_0 \leftarrow \text{ESMFold}(S_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma = 1e^{-4})$
  Sample $x^L_0 \leftarrow \text{HarmonicPrior}(M_{i_{frag}}), \quad \forall frag \in M_i$
  Sample $t \sim \mathcal{U}(0, 1)$
  Concatenate $x_0 = \text{Concat}(x^P_0, x^L_0)$
  Concatenate $x_1 = \text{Concat}(x^P_1, x^L_1)$
  Interpolate $x_t \leftarrow t \cdot x_1 + (1 - t) \cdot x_0$
  Predict $\hat{X}_{h_i} \leftarrow \text{NeuralPLexer}(S_i, M_i, x_t, t)$
  Predict $\hat{B}_i \leftarrow \text{ESDM}_{aff}(S_i, M_i, \text{StopGrad}(\hat{X}_{h_i}))$
  Optimize losses $\mathcal{L}_X := \lambda_X \cdot \text{FAPE}(X_{h_i}, \hat{X}_{h_i}) + \mathcal{L}_B := \lambda_B \cdot \text{MSE}(\hat{B}_i, B_i), \quad \lambda_X = 0.2, \quad \lambda_B = 0.1$
end for

Algorithm 2 Inference

Require: Protein sequences and ligand SMILES strings (S, M)

Ensure: Sampled top-5 heavy-atom structures $\hat{X}$ with confidence scores $\hat{C}$ and binding affinities $\hat{B}$

Sample $x^P_0 \leftarrow \text{ESMFold}(S) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma = 1e^{-4})$
Sample $x^L_0 \leftarrow \text{HarmonicPrior}(M_{frag}), \quad \forall frag \in M$
Concatenate $x_0 = \text{Concat}(x^P_0, x^L_0)$
for $n \leftarrow 0$ to i **do**
  Let $t \leftarrow \frac{n}{i}$ and $s \leftarrow \frac{n+1}{i}$
  Predict $\hat{X} \leftarrow \text{NeuralPLexer}(S, M, x_n, t)$
  if $n = i - 1$ **then**
    Predict $\hat{C} \leftarrow \text{ESDM}_{conf}(S, M, \hat{X})$ # Pre-trained
    Predict $\hat{B} \leftarrow \text{ESDM}_{aff}(S, M, \hat{X})$
    Rank top-5 $\hat{X}$ and $\hat{B}$ using $\hat{C}$
    return $\hat{X}, \hat{C}, \hat{B}$
  end if
  Extract $\hat{x}_1 \leftarrow \text{HeavyAtoms}(\hat{X})$
  Align $x_n \leftarrow \text{RMSDAlign}(x_n, \hat{x}_1)$
  Interpolate $x_{n+1} = clamp(\frac{1-s}{1-t} \cdot \eta) \cdot x_n + clamp((1 - \frac{1-s}{1-t}) \cdot \eta) \cdot \hat{x}_1, \quad \eta = 1$
end for



FlowDock Init. (t=1/40) | **FlowDock Interm.** (t=10/40) | **FlowDock Interm.** (t=20/40) | **FlowDock Interm.** (t=30/40) | **FlowDock Final** (t=40/40)

Fig. 8: Comparison of FlowDock's predicted structure states (w/o hydrogens) for CASP16 superligand pose pharma target L3008.

B. Structure generation example trajectory

To illustrate one of FlowDock's interpretable structure generation trajectories using conditional flow matching, in Figure 8, we report FlowDock's predicted structural states for CASP16 superligand pose pharma target L3008, notably a *multi*-ligand pose target, in evenly spaced increments throughout FlowDock's generation trajectory. In short, we see that FlowDock enables multi-ligand protein complexes to be predicted through concise flow trajectories, yielding early protein and ligand conformational changes following the model's initial binding pocket prediction.

C. CASP16 structure prediction results

In Figure 9, we compare the protein-ligand structure prediction RMSDs of FlowDock and MULTICOM_ligand (Morehead et al., 2025), a top-5 multi-model deep learning prediction method in the CASP16 ligand prediction category, for the 231 superligand pose pharma targets made available during the 16th Critical Assessment of Techniques for Structure Prediction (CASP16). As these results demonstrate, FlowDock, as a standalone deep learning method, achieves competitive structure predictions for many of the new CASP16 ligand targets. Similarly, Figure 10 illustrates that FlowDock and MULTICOM_ligand are approximately tied in terms of their ability to structurally model CASP16's 56 *multi*-ligand protein complexes, further highlighting the broad applicability of FlowDock's structure predictions in diverse drug discovery settings.

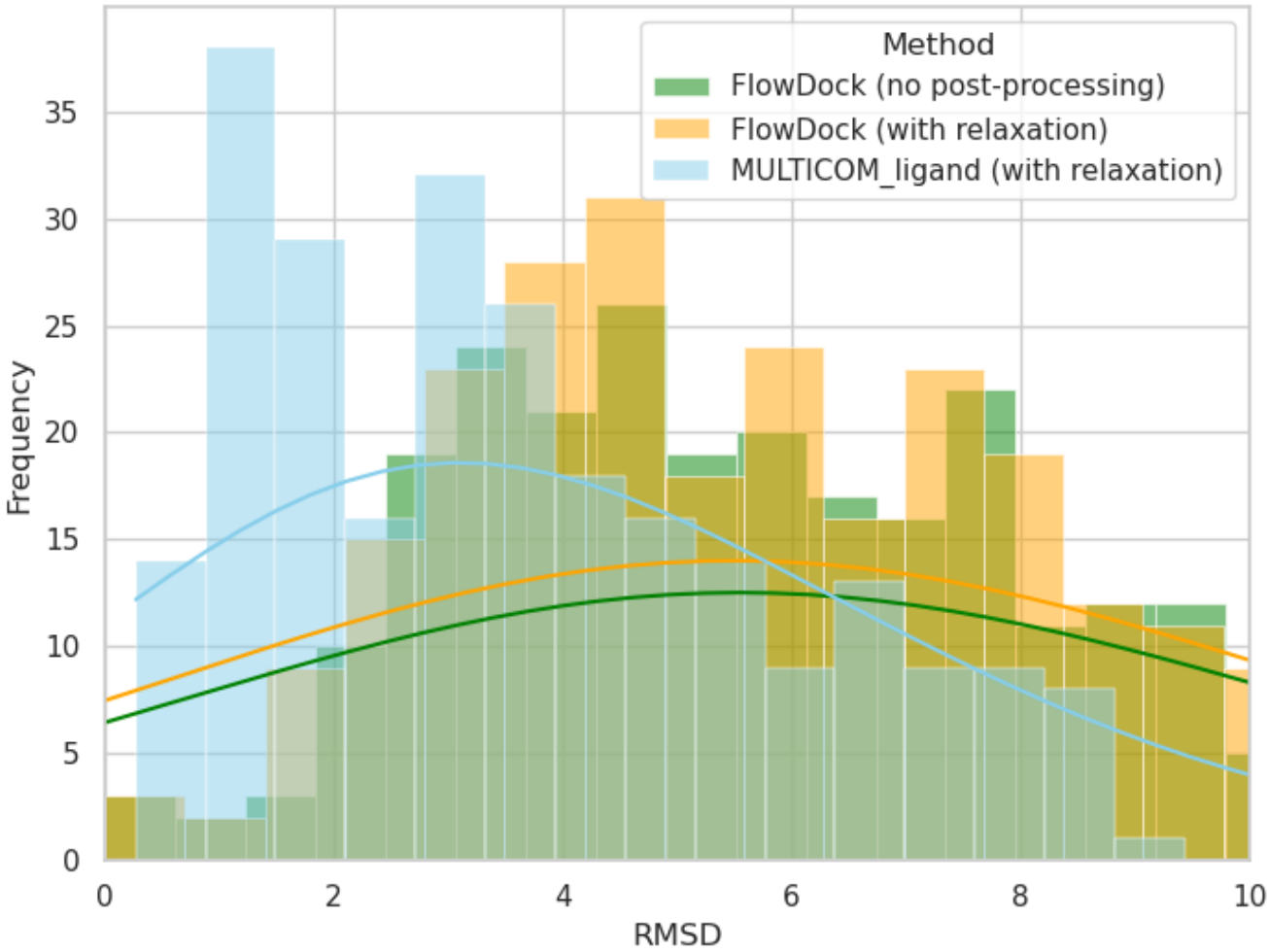


Fig. 9: Comparison of the protein-ligand structure prediction results of FlowDock and the deep learning ensembling method MULTICOM_ligand in terms of their binding pocket-aligned ligand RMSDs for the CASP16 superligand pose pharma targets (n=301).

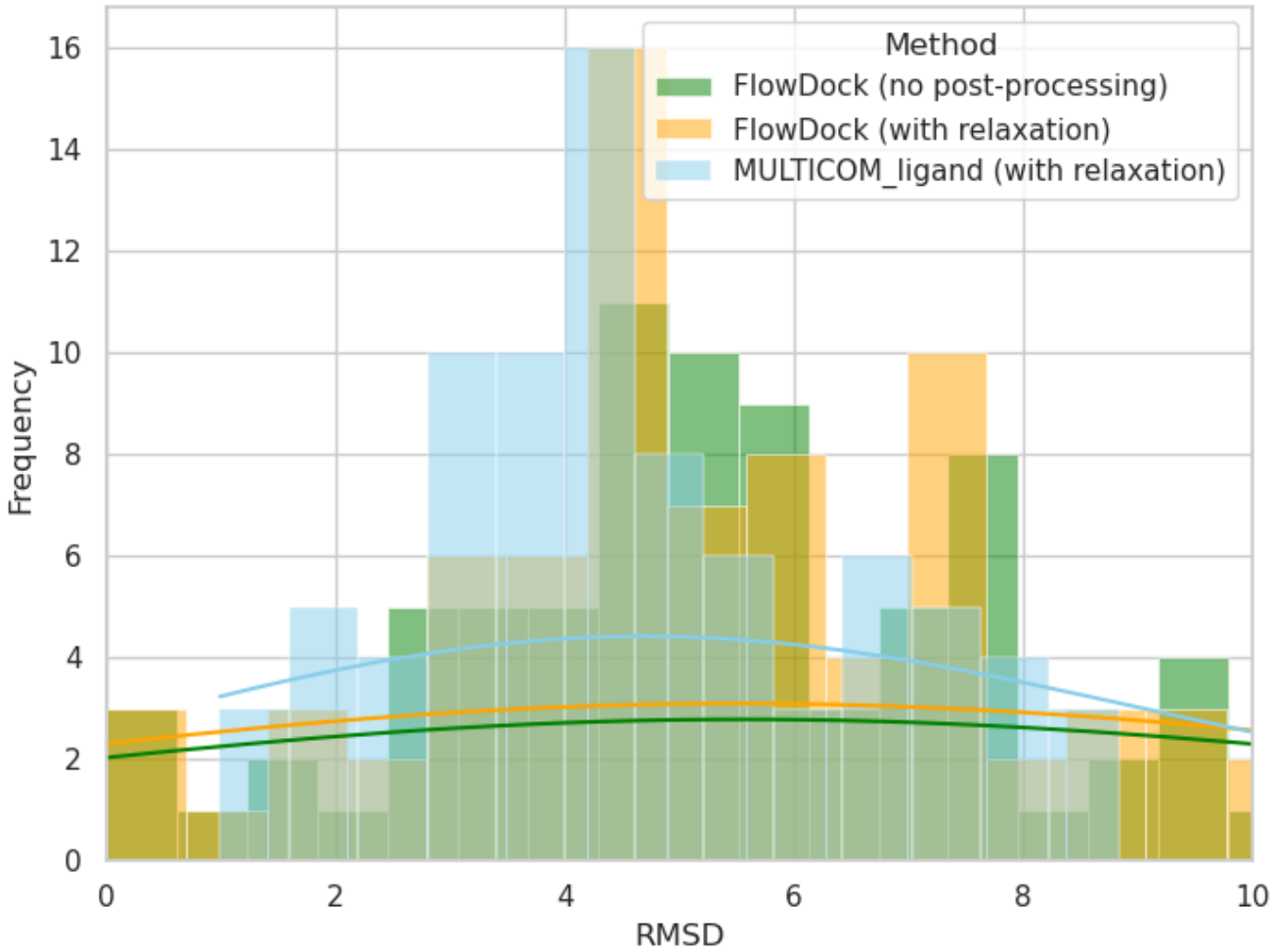


Fig. 10: Comparison of the protein-(multi-)ligand structure prediction results of FlowDock and the deep learning ensembling method MULTICOM_ligand in terms of their binding pocket-aligned ligand RMSDs for the CASP16 superligand pose pharma targets (n=126).

D. PoseBusters Benchmark ligand dissimilarity structure prediction results

To investigate FlowDock's chemical generalization capabilities, in Figure 11, we illustrate the structure prediction performance of FlowDock for chemically dissimilar (Tanimoto similarity < 0.6) ligands associated with the same protein target in the PoseBusters Benchmark dataset. Figure 11 shows that FlowDock's average ligand RMSD of each of these (multi-)ligand protein targets is approximately 2 Å, with a standard deviation around 1 Å, highlighting that its predictions for chemically dissimilar intra-protein ligands are of high average accuracy and demonstrate generalizability with the consistency of FlowDock's average inter-ligand RMSD differences.

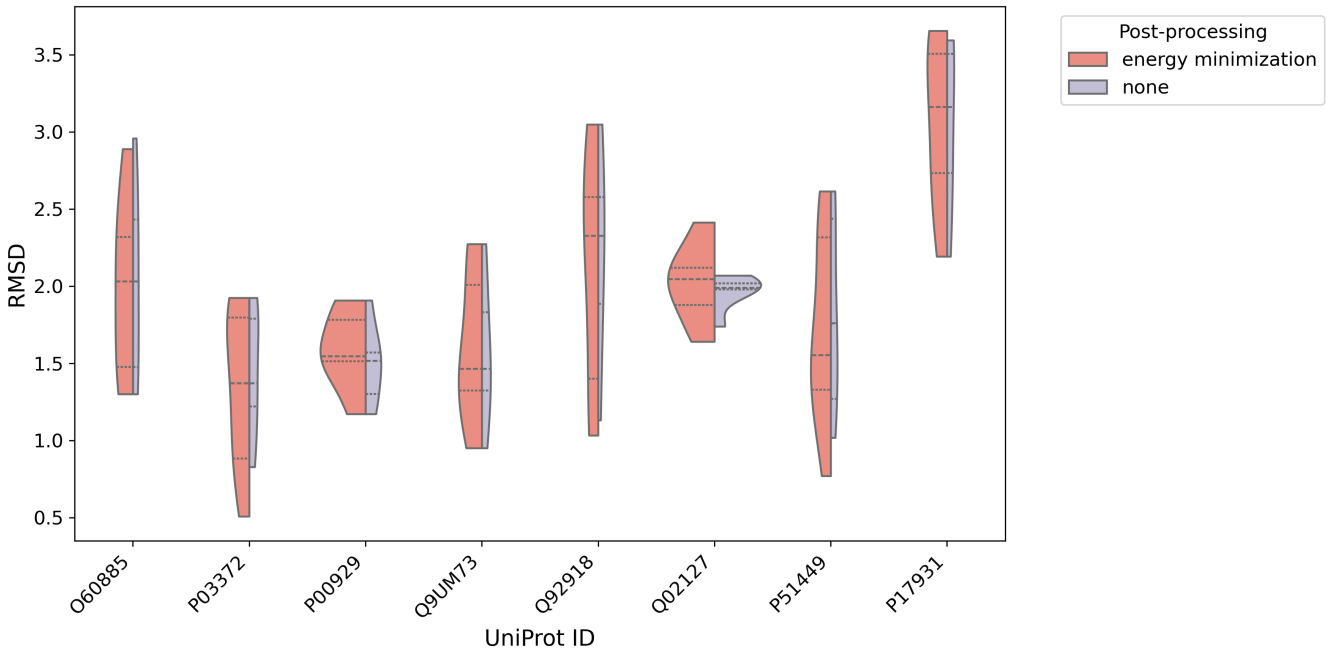


Fig. 11: Analysis of the protein-ligand structure prediction results of FlowDock in terms of its binding pocket-aligned ligand RMSDs for the chemically dissimilar (multi-)ligand PoseBusters Benchmark targets (n=18).