Acta Pharmaceutica Sinica B 2026;16(7):4067−4082

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

FlowDock: A unified flow-based framework for flexible protein-ligand docking and binding affinity prediction

Jing Li, Ruiqiang Lu, Yi Tan, Pengyu Liang, Bo Liu, Shukai Gu, Huanxiang Liu*, Xiaojun Yao*

Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China

Received 22 April 2025; received in revised form 22 July 2025; accepted 11 September 2025

KEY WORDS

Structure-based drug design;
Protein−ligand docking;
Binding affinity prediction;
Machine learning in drug discovery;
Bayesian flow networks;
Deep equivariant generative models;
Latent space optimization

Abstract Accurate prediction of protein−ligand complexes and binding affinity is critical for hit identification and optimization for structure-based drug design. Traditional docking simulates binding processes with searching algorithms guided by energy-scoring functions, which are quite computationally expensive and time-intensive. In contrast, deep learning approaches offer a cost−effective alternative, yet often generate conformations with limited physicochemical validity and fail to account for protein flexibility. To address these pitfalls, we propose FlowDock, a multitask framework enhanced by Bayesian Flow Networks. FlowDock simultaneously generates accurate protein−ligand complex structures and predicts binding affinity while incorporating protein conformational flexibility. By leveraging multimodal intramolecular representations with a deep equivariant generative model, our method iteratively refines complex in latent space, ensuring rapid and stable generation. Benchmark evaluations demonstrate that FlowDock achieves state-of-the-art performance in binding pose prediction, especially physical plausibility, and virtual screening capability, alongside reliable binding affinity predictions. By providing deeper molecular insights into dynamic protein−ligand interactions, FlowDock represents a robust tool for accelerating the rational development of therapeutics.

This article is part of special issue entitled Machine Learning in Drug Discovery.

*Corresponding authors.
E-mail addresses: xjyao@mpu.edu.mo (Xiaojun Yao), hxliu@mpu.edu.mo (Huanxiang Liu).

Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.04.009

1. Introduction

Accurate prediction of protein−ligand binding interactions constitutes the foundation of modern structure-based drug design[1−5], which is critical for virtual screening and therapeutic discovery[6−8]. Elucidating molecular interactions between proteins and potential ligands facilitates the identification and design of compounds to specific targets, thereby optimizing their functions to achieve pharmaceutical effects[9−11]. Over the years, a diverse array of computational methodologies has been devised to address this challenge, ranging from traditional search-and-energy-based methods to cutting-edge deep learning techniques 10. Classical docking methods, which leverage the physicochemical properties of molecules to simulate interactions and predict binding poses with optimal binding energies[9,12,13], have laid the foundation for more advanced, data-driven deep learning frameworks[14−18]. Despite significant progress in predicting either binding poses or binding affinities, there remains a lack of a unified and open framework capable of simultaneously addressing the core tasks required for virtual screening: binding pose generation, binding pose selection, and binding affinity prediction. Such an all-in-one framework is highly sought-after in the early stages of drug discovery and development[19,20].

Moreover, most docking tools require prior knowledge of the binding pocket[21−23], which is often unavailable in many real-world scenarios. From this perspective, blind docking is of more value for drug discovery of new targets without the information on the binding pocket. Most docking methods based on deep learning typically treat proteins as rigid structures, neglecting their inherent flexibility. This violates the fact that proteins also undergo conformational changes during interactions with ligands[24−26], which are critical for accurate binding predictions.

To address these limitations, we propose FlowDock, a novel multi-modal framework boosted by Bayesian Flow Networks, to study protein−ligand blind flexible docking while balancing low computational cost with high accuracy. This multimodal framework integrates the sequence data of residues combined with a coarse-grained 3D protein structure as the initial protein input. For small molecules, it incorporates sequence-level SMILES (Simplified Molecular Input Line Entry System) data[27] and the 3D conformation of the ligand. After initializing these inputs as graph embeddings, the generative model, Bayesian Flow Networks (BFNs)[28] iteratively samples molecular representations to learn the underlying data distribution. The model then predicts the translation, rotation, and side-chain dihedral changes for the protein, as well as the translation, rotation, and torsion angles for the ligand. Simultaneously, FlowDock outputs a confidence score for each binding pose generated from the learned probabilistic distribution. These poses are ranked based on their likelihood of matching the ground-truth binding pose within an acceptable error tolerance. Additionally, FlowDock predicts binding affinities in parallel, further enhancing the utility of the framework for protein-ligand interaction studies and virtual screening tasks. Compared with diffusion-based models, which have recently obtained significant progress in structure prediction tasks, our BFNs-based framework effectively generates more rational conformations by enabling a more stable and consistent sampling, focusing on operations in parameter space rather than directly in data space. This approach achieves a superior balance between efficiency and quality, yielding substantial improvements in both processing speed and predictive accuracy over diffusion-based methods.

In line with prior studies, FlowDock was trained on PDBbind v2020[29], excluding the core set, and evaluated on the CASF 2016 benchmark[30]. To assess its generalizability to novel proteins and ligands, we further tested it with PoseBusters dataset[31] and TrueDecoy set[32]; we also evaluated its screening power with the DEKOIS 2.0 dataset[33]. Experimental results demonstrate that FlowDock achieves state-of-the-art performance in binding pose generation and virtual screening across various metrics, showing comparable binding affinity prediction performance to complex-based methods[34−39] while surpassing all complex-free methods[40−44]. Experimental results also reveal significant improvements in generation stability and reduced computational costs.

To the best of our knowledge, FlowDock represents the first unified, multi-modal, multi-task flow-based framework designed to generate protein−ligand complex structure alongside the corresponding binding affinity for structure-based drug design. Unlike traditional approaches reliant on predefined binding sites, FlowDock employs flexible blind docking, allowing for a more comprehensive exploration of protein−ligand interactions. We anticipate that FlowDock has the potential to facilitate structure-based virtual screening tasks and hit compound identifications in drug development.

2. Materials and methods

2.1. Datasets curation

We downloaded the PDBbind database to train and evaluate FlowDock. This dataset contains a total of 19,443 protein−ligand complex structures, each with an experimentally determined binding affinity. To ensure a fair comparison with baseline methods, we followed the same dataset splitting protocol[15,44,45]. A general set excluding the complexes that exhibited protein sequence similarity above 40% and ligand fingerprint similarity above 80% with any complex in the test dataset (18,404 complexes) is used for training, while the core set CASF 2016 containing 285 complexes is used for testing. Our scaffold-based splitting strategy (ligand fingerprint similarity $\leq$80%) evaluates the model's ability to generalize to novel chemotypes, providing a complementary perspective to the commonly employed time-based splitting approach, which assesses temporal generalization.

Furthermore, to evaluate the generality of our framework, we used another dataset, PoseBusters[31], as the benchmark and compared it with the baseline methods. The PoseBusters set encompasses a diverse array of high-caliber, recent protein−ligand complexes characterized by drug-like molecules. With 308 distinct complexes, inclusive of unique proteins and ligands released since 2021, it ensures no overlap with the complexes found in the PDBbind v2020 dataset.

Moreover, DEKOIS 2.0 provides 81 structurally diverse protein sets spanning a broad range of protein classes, each accompanied by 1200 carefully selected decoys and 40 true binding ligands. The data from this dataset is guaranteed to be unseen for FlowDock training to prevent data leakage[46]. We use it to evaluate the screening performance for 81 diverse protein targets. Furthermore, to assess the ultra-VS performance, we use ChemDiv, which contains over 1.5 million compounds, to compare the screening speed.

Nevertheless, TrueDecoy set is a recently published dataset for benchmarking AI-powered deep learning tools[47]. It contains a

wide range of 147 targets, including kinases, GPCRs, peptidases, etc. With the decoys being all truly experimentally valid, this provides another perspective for evaluating docking tools.

In summary, the exclusive core-set-exclusive general set from PDBbind v2020 served as the training dataset, while CASF 2016, PoseBusters, DEKOIS 2.0, and TrueDecoy sets were utilized to evaluate the model.

2.2. Problem formulation

To investigate protein-ligand interactions, we primarily focus on predicting both binding poses and corresponding binding affinities. Two main challenges arise in binding pose prediction. First, in real-world applications, prior knowledge of the protein binding pocket is often unavailable, making blind global docking more pertinent than docking with a known pocket. Our objective is to predict the binding pose of a protein−ligand pair using only the apo structures of the protein predicted by AlphaFold2 and the ligand initialized with RDKit. Second, while most approaches to binding pose prediction concentrate solely on ligand conformational changes, they often overlook the fact that proteins also undergo conformational changes during the docking process. In summary, our work emphasizes the importance of capturing the conformational dynamics of both the ligand and the protein in unconditioned pockets.

To address these challenges, we formulate the problem as follows. Given a pair of proteins and ligand (P, L), where P denotes the protein and L denotes the ligand, our objective is as follows:

(1) Predict the binding pose, which involves not only changes in ligands: translation t^l, rotation r^l, and torsion τ^l; but also changes in proteins: the dihedral angles of side chains τ^p, translation t^p and rotation r^p of backbones.
(2) Rank generated binding pose based on the output confidence score C, which demonstrates how reliable the predicted structure is within the preset tolerance error;
(3) Predict the binding affinity A for each of the generated samples.

To model the joint conformational landscape of flexible protein−ligand systems under the constraints of 3D geometric invariance, we adopt the Bayesian Flow Networks (BFNs) framework, designed to operate over 3D molecular geometries in a manner that is invariant to rigid-body transformations, including rotations and translations under the SE (3) group. Unlike discrete diffusion models that struggle with gradient discontinuities in torsional space, BFNs operate in parameter space to smoothly evolve joint probability distributions over ligand torsion angles τ^l and protein side-chain dihedrals τ^p. This framework naturally accommodates the correlated flexibility of bimolecular systems through factorized distributions that concurrently represent continuous variables like coordinates. This makes them particularly well-suited for tasks involving spatially flexible molecular systems. The generative process can be interpreted as a flow of information from noise to structured data, where the latent variables are designed to change smoothly and progressively capture more structural detail. This formulation enables the BFNs to jointly model the conformational changes of both the ligand and the protein. During inference, the BFNs generate candidate binding poses by sampling from the learned latent space, with each sample representing a plausible protein−ligand complex. These samples are then scored and ranked to identify the most likely binding conformation and estimate the associated binding affinity.

2.3. Development of the FlowDock architecture for protein-ligand complex generation and binding affinity prediction

FlowDock structure consists of two stages: the representation learning stage and the docking prediction stage. On the representation learning stage (Fig. 1A), FlowDock takes both the sequence-level and structure-level data of molecules. On the one hand, for proteins, amino acid sequences are processed with a protein language model (PLM), ESM2[48], to get corresponding sequence embeddings. Meanwhile, the coarse-grained 3-dimensional protein structures are represented *via* graph neural networks (GNNs). On the other hand, for small molecules, we use RDKit[49] to handle SMILES (Simplified Molecular Input Line Entry System) data to obtain a seed conformation. RDKit is employed to generate initial three-dimensional ligand conformations from their corresponding two-dimensional molecular graphs, in our case, the input SMILES strings. This is accomplished using RDKit's ETKDG algorithm, which produces chemically and physically plausible conformers by integrating knowledge-based torsion preferences with distance geometry. We also use molecular graphs to encode structural attributes, including bond types, atomic number, chirality, degree, bond length, and formal charge. The graph representations combined with the sequence embeddings are concatenated together as a complete initialization of molecules and are put into the docking prediction stage, as shown in Fig. 1B, executing core computational tasks. The input embeddings are first fed into an MLP module to obtain a joint representation for proteins and ligands. Then, a triangular attention module, inspired by AlphaFold2[50], processes the matrix to learn important weights of nodes to capture inter-molecular interactions. Next, an SE (3)-Equivariant module[51] is added to guarantee the equivariance of geometric properties. Subsequently, the BFN Module generates the predicted translation, rotation, and torsion updates for both proteins and ligands. Specifically, a simple prior (Gaussian) distribution serves as the input distribution. At each generation step, the parameters of this input distribution are passed to a neural network, which generates an output distribution that integrates all messages across variables of data. At the same time, a sender distribution is created by adding noise to the sampled data, and a receiver distribution is then formed by combining the output distribution with the same noise used in the sender distribution.

Using Bayesian updates, the receiver distribution is employed to refine the input distribution, progressively improving learning of underlying data distributions. After the Bayesian updates, the refined input distribution is passed through an SE (3)-Equivariant module. The module captures geometric relationships in graph-structured data, allowing the model to learn spatial interactions and further refine node embeddings. The embeddings are then integrated back into the output distribution. This iterative process is repeated for multiple steps, and after sufficient refinement, the output distribution represents the best estimate of the original data distributions, allowing for accurate sample generation (Fig. 1C).

Unlike diffusion models, which sample from the original data space directly[52], the BFN module samples in the differentiable parameter space, resulting in a continuous training procedure even for discrete data like formal charges. A graphical model has been displayed to clarify the procedure. Diffusion-based models exhibit

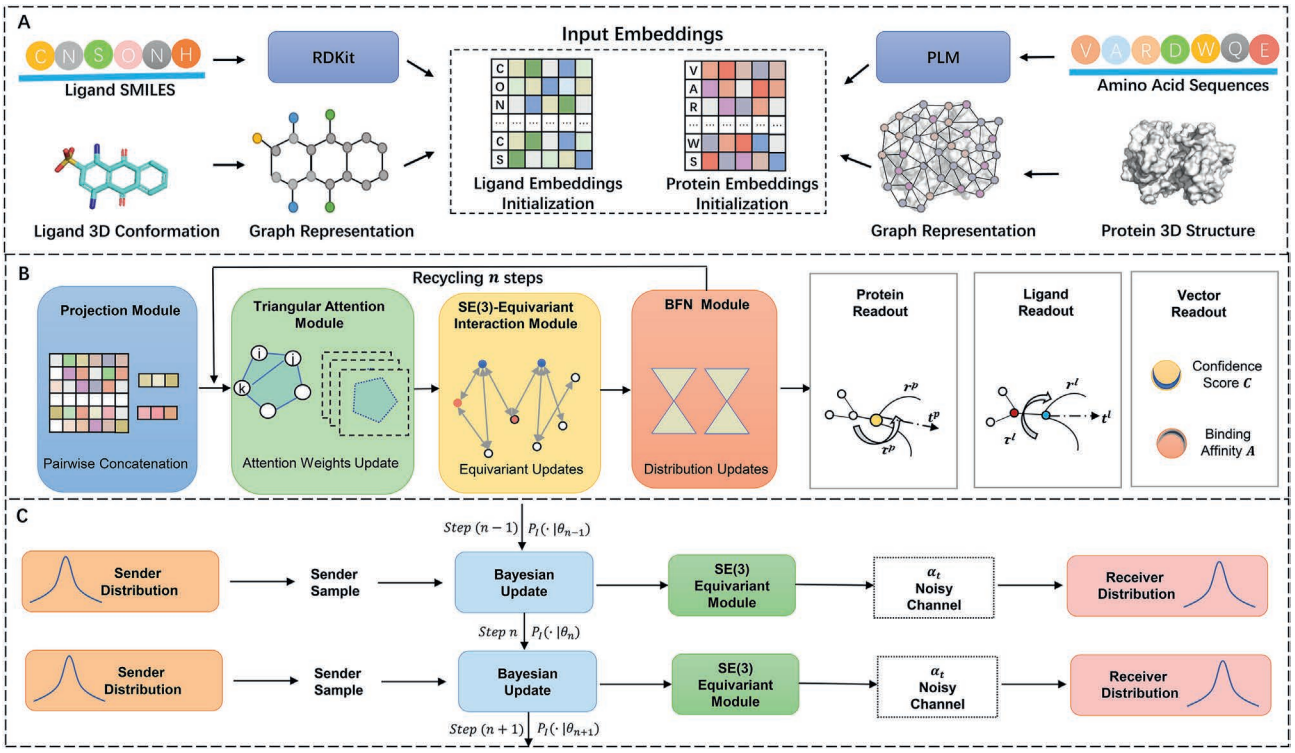


Figure 1 Overview of FlowDock architecture. (A) Representation learning. Multimodal ligand and protein data are taken as inputs to get sequence embeddings and graph representations. (B) Docking Prediction. The input embeddings are firstly mapped by a projection module, and then processed by a triangular attention module. To maintain the equivariance of modules, we utilize a SE (3)-equivariant module. Lastly, the BFN module is used to generate the binding pose. Additionally, we train a confidence model to rank the quality of generated conformation and also output corresponding binding affinity. (C) Structure of BFN Module. The BFN Module learns data distribution by recursive sampling in the latent space.

noise sensitivity as perturbing the native complex with multiple magnitudes of noise and denoising to learn distributions. Our model addresses this issue by sampling from the parameters of the distribution instead of the data itself, which guarantees the complete continuity of the transmission process and the stability and consistency of generation for various data types. In this way, FlowDock strikes a balance between generation quality and efficacy.

2.4. *Multimodal intramolecular representations of proteins and ligands*

To extract features of proteins, we use the multimodal representations of proteins. For the sequence representation of protein, we take the sequence of amino acids as input, and then process it with protein large language model (PLM), ESM2[48], to get the corresponding embeddings $\mathscr{S}^p$ that contain the whole protein information, including its structure and function. For the geometric representation of a protein, we model the 3-dimensional pocket in a coarse-grained pattern of backbone plus residues. The protein graph was represented by $\mathscr{G}^p = (\mathscr{V}^p, E^p)$, where each node $v_i^p \in \mathscr{V}^p$ refers to the residue located at C_α position. $\varepsilon_i^p \in E^p$ denotes the edge, denoting the bonds between residues. The node attributes contain encoding of amino acid types and side-chain dihedral angles computed from at most five rotatable χ angles and two symmetric χ angles of each amino acid. For the uniqueness and consistency of the side-chain angles, we uniformly process the dihedral angles with:

$$[\max(\chi_1, alt\chi_1), \min(\chi_1, alt\chi_1), \max(\chi_2, alt\chi_2), \min(\chi_2, alt\chi_2), \chi_3, \chi_4, \chi_5] \quad (1)$$

instead of $[\chi_1, \chi_2, \chi_3, \chi_4, \chi_5]$. Furthermore, the orientation of the backbones are determined by the positions of atoms $\{C_\alpha, C, N\}$, which are represented by two vectors $\frac{x_N - x_{C_\alpha}}{||x_N - x_{C_\alpha}||}$ and $\frac{x_C - x_{C_\alpha}}{||x_C - x_{C_\alpha}||}$ for every node. For edge attributes, we incorporate the bonds types and length. In this way, we construct the protein graph containing the important spatial information that facilitates the model to derive the relative positions of all heavy atoms of proteins. In short, the representations for protein initializations concatenate the sequence embedding and the graph embedding, represented as Eq. (2):

$$P^0 = (\mathscr{S}^p, \mathscr{G}^p) \quad (2)$$

Here, $\mathscr{S}^p$ denotes the sequence embeddings, and $\mathscr{G}^p$ denotes the 3D structural embeddings for proteins.

For small molecules, we also extract the features in two patterns. On the one hand, we take the 1D SMILES sequence $\mathscr{S}^l$ as input and process it with RDKit. On the other hand, we also use the molecular graph $\mathscr{G}^l = (\mathscr{V}^l, E^l)$ to represent the 3D structure of ligands abstracted from complex structures. Nodes denote atoms and edges denote covalent bonds. In this ligand graph $\mathscr{G}^l$, the node attributes contain the atom types, degree, chirality, torsion angles, and number of formal charges, which are concatenated with sinusoidal embeddings. The edge features contain bond types and length, processed with radial basis embeddings. In summary, the initial representations for ligands also connect the

sequence embedding $\mathscr{S}^l$ and the graph embedding $\mathscr{G}^l$, represented as Eq. (3):

$$L^0 = \left(\mathscr{S}^l, \mathscr{G}^l\right) \tag{3}$$

Together, we acquire the initial input embeddings of a protein−ligand pair $\left(P^0, L^0\right)$.

2.5. *Generation via Bayesian Flow Networks*

The initial embeddings of ligands and proteins are processed through an MLP (multilayer perceptron) layer, yielding a unified representation of the protein−ligand pair. This unified pair representation is then analyzed by a triangular self-attention module to examine the weights of each node (Eq. (4)).

$$G^0 = \textbf{TrianAtten}\left(\textbf{LinProj}\left(P^0, L^0\right)\right) \tag{4}$$

When generating the rational binding pose of ligands and proteins, SE (3) equivariance must be guaranteed. Therefore, we use an EGNN module to update the features (Eq. (5)).

$$\widehat{G}^0 = \textbf{EGNN}\left(G^0\right) \tag{5}$$

Next, we process the embeddings with a BFN module, which leverages Bayesian inference to learn the distribution of given data through several interconnected distributions. For simplicity, we take x to denote the initial input of the BFN module as an example. The *input distribution* $p_I(\boldsymbol{x}|\theta)$, set as factorized isotropic Gaussian distributions $\mathscr{N}\left(\cdot\left|\mu, \rho^{-1}\boldsymbol{I}\right.\right)$ with parameters $\theta := \{\mu, \rho\}$, serves as a starting guess of data. And the *sender distribution* p_S is also set as a Gaussian distribution, allowing the model to sample different variations of the input with a specific known noise level α (Eq. (6)):

$$p_S(y|x;\alpha\mathbf{I}) = \mathscr{N}\left(\mathrm{x}|\alpha^{-1}\mathbf{I}\right) \tag{6}$$

As illustrated in Fig. 1C, the Bayesian update function is utilized to update the parameters (Eq. (7)):

$$\mathrm{h}(\{\mu_{i-1}, \rho_{i-1}\}, y, \alpha) = \{\mu_{\mathrm{i}}, \rho_i\}, \text{where } \rho_i = \rho_{i-1} + \alpha, \mu_i = \frac{\mu_{i-1}\rho_{i-1} + y\alpha}{\rho_i}. \tag{7}$$

The receiver distribution subsequently combines information from the output distribution p_O and sender distribution p_S to assess the reconstruction of noisy data (Eq. (10)):

$$p_R(y|\theta,\alpha) = \mathop{\mathbb{E}}_{p_O(x'|\theta)}\left[p_S(y|x';\alpha)\right] \tag{10}$$

Finally, the variational lower bound of likelihood is optimized as proven in GeoBFN[53]. For learning the probability distribution p_θ over given x, a series of noisy versions $< y_1, \cdots, y_n >$ of x are introduced as latent variables (Eq. (11)):

$$\mathscr{L}_{VLB}(x) = \mathop{\mathbb{E}}_{p_\phi\left(\theta_0^x, \cdots, \theta_n^x\right)}\left[\sum_{i=1}^{n} D_{KL}\left(p_S(\cdot|x;\alpha_i)\middle\|p_R\left(\cdot\middle|\theta_{i-1}^x;\alpha_i\right)\right) - \log p_\phi\left(x\middle|\theta_n^x\right)\right] \tag{11}$$

With the learned distribution being *Bayesian update distribution* factorized (Eq. (12)):

$$p_\phi\left(\theta_0^x, \cdots, \theta_n^x\right) = \prod_{i=1}^{n} p_U(\theta_i|\theta_{i-1}, x;\alpha_i).$$
$$p_R\left(\cdot\middle|\theta_{i-1}^x;\alpha_i\right) = \mathop{\mathbb{E}}_{p_O\left(\theta_{i-1}^x;\phi\right)}\left[p_S\left(y_i\middle|x_i';\alpha_i\right)\right], p_\phi\left(x\middle|\theta_n^x\right) = p_O\left(x|\theta_n^x,\phi\right) \tag{12}$$

Unlike diffusion models operating on a noisy version of data samples, the BFN module acts upon the distribution parameters of data. And this inherent scheme guarantees the generative process is completely continuous and differentiable, even when dealing with discretized data like residue and atom types.

In particular, during each step of the graph propagation process, the ligand and protein graphs undergo both intramolecular and intermolecular interaction updates. For intramolecular interactions, the representation of each ligand atom is refined based on its neighborhood ligand atoms within a 6 Å radius. Similarly, each amino acid in the protein graph is updated based on its neighboring residues within a 12 Å radius, with a maximum of 30 neighbors per residue enforced to optimize computational efficiency during training. Once the overall graph is established, the message between nodes is updated by spherical harmonics and the tensor product. More specifically, the updated procedure follows (Eq. (13)).

$$\boldsymbol{h}_a \longleftarrow \boldsymbol{h}_a \bigoplus_{t\in\{\mathscr{C},r\}} \mathbf{BN}^{(t_a,t)}\left(\frac{1}{\left|\mathscr{N}_a^{(t)}\right|}\sum_{b\in\mathscr{N}_a^{(t)}} Y(\widehat{r}_{ab})\bigotimes_{\psi_{ab}}\boldsymbol{h}_b\right) \text{ with } \psi_{ab} = \Psi^{(t_a,t)}\left(e_{ab}, \boldsymbol{h}_a^0, \boldsymbol{h}_b^0\right) \tag{13}$$

And the corresponding Bayesian update distribution is derived as (Eq. (8)):

$$p_U(\theta_i|\theta_{i-1}, x;\alpha) = \mathscr{N}\left(\mu_i\middle|\frac{\alpha x + \mu_{i-1}\rho_{i-1}}{\rho_i}, \frac{\alpha}{\rho_i^2}I\right) \tag{8}$$

Simultaneously, a neural network Ψ takes θ and time t as input to generate the *output distribution* (Eq. (9)):

$$p_O(y|\theta,t) = \delta(x - \Psi(\theta, \mathrm{t})) \tag{9}$$

where δ denotes Dirac delta distribution, and the accuracy scheduler is defined as $\beta(t) = \int_{t'=0}^{t}\alpha(t')dt', t\in[0,1]$.

Here, $\boldsymbol{h}_a$ denotes the node embeddings belong to type a (either being a residue or atom), $\mathscr{N}_a^{(t)} = \{b|(a,b)\,\epsilon\,\mathscr{E}_{t_a,t}\}$ the neighbors of a of type t, $\widehat{r}_{ab}$ the unit vector in the direction of node a to b, Y being the spherical harmonics and BN the batch normalization layer. All weights are included in the MLP module Ψ to compute the message passing along the scalar features $\boldsymbol{h}_a^0, \boldsymbol{h}_b^0$ and edge embeddings e_{ab}.

Following the final interaction layer, the node representations are utilized to generate the output predictions. As illustrated in Fig. 1B, their final output consists of four parts. For generating the pLDDT, binding affinity, translation, and rotation of ligands, a

convolution of each ligand atom node with the geometric center of the ligand is employed (Eq. (14)):

$$\boldsymbol{v} \leftarrow \frac{1}{\mathscr{Y}^{\mathscr{A}}} \sum_{a \epsilon \mathscr{Y}^{\mathscr{A}}} Y(\hat{r}_{oa}) \bigotimes_{\psi_{oa}} \mathbf{h}_a \text{ with } \psi_{oa} = \Psi\left(e_{oa}, \mathbf{h}_a^0\right) \tag{14}$$

where $\hat{r}_{oa}$ denotes the unit vector from center of ligand to node a, e_{oa} denotes the edge embedding of center and node a. The output $\boldsymbol{v} = [v_{scalar}, \boldsymbol{v}_{odd}, \boldsymbol{v}_{even}]$ consists of 36 even scalars, 1 odd vector and 1 even vector, prepared for further processing.

2.6. Confidence score and binding affinity regression

To effectively train a pLDDT model for molecular docking poses, we employ a systematic approach involving the generation and evaluation of pose candidates using a trained generative model.

For each training instance, a trained BFN model is used to generate a comprehensive range of candidate docking poses, exploring a broad conformational spectrum and capturing multiple plausible binding scenarios for target protein−compound interactions. Every generated pose is meticulously evaluated to calculate pLDDT scores for specific regions or residues within the binding interface. These scores, ranging from 0 to 1, indicate the local accuracy of docking conformations, with higher values reflecting greater confidence. They are derived from the deviation of each candidate pose from a high-confidence reference binding pose, focusing on per-residue accuracy. The training process employs a regression approach, using a loss function MSE (Mean Squared Error) to optimize the model's pLDDT predictions. During the inference phase, the BFN module generates N candidate poses simultaneously, ensuring comprehensive examination of potential docking configurations. Each pose is evaluated by pLDDT, which assigns scores indicative of the pose's local accuracy.

The output $\boldsymbol{v} = [v_{scalar}, \boldsymbol{v}_{odd}, \boldsymbol{v}_{even}]$ consists of 36 even scalars, 1 odd vector and 1 even vector. The scalars are used for predicting the confidence score C and negative logarithm of the binding affinity A by the following Eqs. (15) and (16):

$$C = MLP(v_{scalar}[:18]) \tag{15}$$

$$A = \text{clip}\left(\frac{MLP}{C+\varepsilon}, min = 0, max = 15\right) \tag{16}$$

The binding affinity regression is also trained using MSE. Poses are ranked based on their computed confidence scores, thereby enabling the identification and prioritization of the most promising poses for further experimental validation or subsequent computational analysis.

$\boldsymbol{v}_{odd}$ is used to predict the ligand translation t^l and $\boldsymbol{v}_{even}$ is used to predict ligand rotation r^l (Eqs. (17) and (18)):

$$t^l = \frac{\boldsymbol{v}_{odd}}{||\boldsymbol{v}_{odd}|| + \varepsilon} \cdot MLP(||\boldsymbol{v}_{odd}||, s_t) \tag{17}$$

$$r^l = \frac{\boldsymbol{v}_{even}}{||\boldsymbol{v}_{even}|| + \varepsilon} \cdot MLP(||\boldsymbol{v}_{even}||, s_t) \tag{18}$$

where s_t is the sinusoidal embeddings of schedule time of BFN and $\varepsilon = 10^{-10}$ is added in case the nominator vanishes. Moreover, a scalar torsion update for each rotatable bond b of ligand is predicted by a convolution (Eq. (19)):

$$\tau_b^l = MLP\left(\frac{1}{|\mathscr{N}_b|} \sum_{a \in \mathscr{N}_b} Y(\hat{r}_{oa}) \bigotimes Y^2(r_b) \bigotimes_{\gamma_{oa}} h_a\right) \quad \text{with } \gamma_{oa} = U\left(e_{oa}, \mathbf{h}_a^0, \mathbf{h}_a^0 + \mathbf{h}_b^0\right) \tag{19}$$

For generating the conformation changes of protein, we require to output the translation, rotation, and torsion of side-chains, denoted as t_i^p, r_i^p and τ_i^p for the i-th residue, respectively.

These are obtained from the representations h_i of i-th residue after the final interaction layer (Eqs. (20)−(22)):

$$\tau_i^p = MLP\left(\boldsymbol{h}_{i,odd}, \boldsymbol{h}_{i,even}\right) \tag{20}$$

$$t_i^p = \frac{\widehat{\boldsymbol{h}}_{i,odd}}{||\widehat{\boldsymbol{h}}_{i,odd}|| + \varepsilon} \cdot MLP\left(\left\|\widehat{\boldsymbol{h}}_{i,odd}\right\|, s_t\right) \tag{21}$$

And

$$rot_i^p = \frac{\widehat{\boldsymbol{h}}_{i,even}}{||\widehat{\boldsymbol{h}}_{i,even}|| + \varepsilon} \cdot MLP\left(\left\|\widehat{\boldsymbol{h}}_{i,even}\right\|, s_t\right) \quad \text{with } \widehat{\mathbf{h}}_i = \frac{1}{|\mathscr{N}_i|} \sum_{j \epsilon \mathscr{N}_i} h_i^j \tag{22}$$

where $\mathscr{N}_i$ denotes the neighborhood of i-th residue, and τ_i^P is a five-dimensional scalar vector as defined in Eq. (1).

2.7. Train and inference procedure

2.7.1. Training procedure

The network is trained with a loss that consists of eight components. The total loss is defined as follows Eq. (23):

$$\mathscr{L} = \frac{1}{6}\mathscr{L}_t^p + \frac{1}{6}\mathscr{L}_r^p + \frac{1}{6}\mathscr{L}_\tau^p + \frac{1}{6}\mathscr{L}_t^l + \frac{1}{6}\mathscr{L}_r^l + \frac{1}{6}\mathscr{L}_\tau^l + 0.1\mathscr{L}_A + 0.9\mathscr{L}_C, \tag{23}$$

where $\mathscr{L}_t^p$, $\mathscr{L}_r^p$, $\mathscr{L}_\tau^p$ denote the loss for translation, rotation, and torsion of dihedral angles of residues for proteins. $\mathscr{L}_t^l$, $\mathscr{L}_r^l$ and $\mathscr{L}_\tau^l$ are the losses for heavy atoms of compounds. The translation loss is computed after the optimization process with the Kabsch algorithm. The rotation loss is computed in consideration of symmetry to take the minimum of the orientation. The torsion loss is computed using the cosine of the angle difference. $\mathscr{L}_A$ denote the loss for binding affinity, and $\mathscr{L}_C$ denote the pLDDT loss calculated by RMSE (more details of the calculation are provided in "Evaluation strategies"). Hyperparameters used during training are listed in Supporting Information Table S1, and the complete training procedure is outlined in Supporting Information Algorithm S1.

2.7.2. Inference procedure

Our framework necessitates multi-modal data input, and these diverse modalities complement each other to achieve optimal performance as demonstrated in the Results section. However, if not given the multimodality of data, the structural data is of more importance than sequence data. During the inference phase, we begin with ligand conformations generated by RDKit and protein structural predictions from AlphaFold2 as the initial configuration, with SMILES sequence or residue sequence being optional. The conformation of this complex undergoes an iterative update process spanning 20 steps. To mitigate the risk of the final conformation becoming trapped in a local minimum, a small random perturbation is introduced to the refined ligand pose at each step. For each ligand-protein pair, we conduct 40 sampling iterations and rank the resultant binding conformations according to the pLDDT values. The detailed inference methodology is described in Supporting Information Algorithm S2.

2.8. Evaluation strategies

With the framework established, we evaluated FlowDock on the test set CASF 2016 and then on two additional benchmarks, PoseBusters and DEKOIS 2.0, which we don't use for training but only for testing to guarantee the data is unseen when training the model. Meanwhile, we independently evaluate three tasks: binding pose generation, binding affinity regression, and screening power. The following metrics pertain to each task.

2.8.1. Evaluation metrics for binding pose quality

Most research studies utilize the Root Mean Square Deviation (RMSD) of ligands to assess the predictive quality of ligand conformations without taking the conformation of proteins into account. However, the measurement of protein conformation is also of importance. Thus, we evaluate the quality of the side chain conformations using pLDDT, which assesses the reliability of predicted atomic positions within a protein model. Specifically, it provides a per-residue measure of confidence in the local geometry of the predicted structure, with higher pLDDT values indicating greater confidence. Here are details of how these metrics are calculated.

2.8.1.1. Ligand RMSD. To evaluate the generated complexes, we compute the heavy-atom RMSD between the predicted and the crystal ligand atoms when the protein structures are aligned. We consider RMSD of ligand corrected for symmetry. The precise calculation formula is given below as Eq. (24):

$$Ligand\ RMSD = \underset{X^{isom} \sim isom(X^{gt})}{\operatorname{argmin}} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X^{isom}(i) - X^{pred}(i)\right)^2} \tag{24}$$

In this context, N denotes the quantity of heavy atoms present in the ligand, while *isom* refers to the various isomers associated with the molecular graph of the ligand. After calculating the RMSD, the cumulative plot of ligand RMSD distribution is presented for comparison with baselines. The higher the proportion of the ligand RMSD below a certain threshold is, the better the model performance is.

2.8.1.2. Success rate. A widely accepted benchmark for successful docking is achieving a ligand RMSD less than 2 Å. The success rate is calculated by dividing the number of successful docking pairs by the total number of samples in the test set.

2.8.1.3. pLDDT. Predicted Local Distance Difference Test (pLDDT) assesses the confidence in local structural predictions by estimating how well the prediction agrees with experimental structures. The pLDDT score measures the proportion of conserved interatomic distances against four specified tolerance thresholds: 0.5, 1, 2, and 4 Å. This score is calculated using the predicted histogram densities associated with each entity (Eq. (25)).

$$\begin{aligned} pLDDT[i] = {} & 1.0 * p_{0-0.5}[i] + 0.75 * p_{0.5-1}[i] + 0.5 * p_{1-2}[i] \\ & + 0.25 * p_{2-4}[i] \end{aligned} \tag{25}$$

2.8.1.4. Physical validity. The structural validity of predicted molecular binding poses was evaluated using PoseBusters, a computational tool leveraging the RDKit library to perform standardized quality evaluations on docked structures. These assessments encompass three primary criteria: (a) chemical validity and consistency, ensuring ligand conformations adhere to stereochemical rules including chirality, bond stereochemistry and chemical feasibility; (b) intramolecular validity, assessing geometric rationality through bond lengths, angles, internal steric clashes, and energy profiles; and (c) intermolecular validity, scrutinizing steric and electronic complementarity between ligand and protein binding sites. Conformations satisfying all three criteria were deemed physically valid.

2.8.2. Evaluation metrics for virtual screening

The enrichment factor[54–56] computed from the top-$x\%$ samples are denoted by $\mathrm{EF}_{x\%}$ $(\mathrm{x} = 0.1, 0.5, 1, 5)$ (Eq. (26)):

$$\mathrm{EF}_{x\%} = \frac{N_{active,top-x\%}}{N_{active} * x\%} \tag{26}$$

Moreover, area under the receiver operator characteristic curve (ROC_AUC), area under the precision recall curve (PR_AUC), and Boltzmann-Enhanced discrimination of the receiver operating characteristic setting $\alpha = 80.5\%$ ($\mathrm{BED\_ROC}_{\alpha=80.5\%}$) are also used for evaluation of screening performance[54–56].

2.8.3. Evaluation metrics for binding affinity accuracy

In drug discovery, both the screening process and the reverse screening process play essential roles[57]. The efficacy of these approaches hinges upon the rapid and precise prediction of binding affinities, a critical metric quantifying the interaction strength between a protein and a ligand, providing valuable insights into molecular recognition and the underlying mechanisms of protein-ligand interactions[58]. The following are the metrics that we use.

2.8.3.1. MAE. The evaluation uses MAE (Mean Absolute Error) for prediction accuracy. MAE is commonly used for regression-based tasks, which measure the average absolute difference between the predicted binding affinity A_i^{pred} and the true values A_i^{true}. The formula is as follows (Eq. (27)):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|A_i^{true} - A_i^{pred}\right| \tag{27}$$

2.8.3.2. RMSE. Root Mean Square Error (RMSE, Eq. (28)) is used to be more sensitive to significant errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(A_i^{true} - A_i^{pred}\right)^2} \tag{28}$$

2.8.3.3. Spearman's correlation. This evaluates the correlation of predicted with true experimental affinities (Eq. (29)).

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}\left(A_i^{true} - A_i^{pred}\right)^2}{n(n^2-1)} \tag{29}$$

2.9. Data availability

Raw data for training FlowDock were collected from PDBbind v2020 (http://www.pdbbind.org.cn). The PoseBusters dataset used was downloaded from https://github.com/maabuu/posebusters. The DEKOIS 2.0 was downloaded from http://www.dekois.com.

The TrueDecoy set was downloaded from https://zenodo.org/records/13684010. Source data are provided with this paper.

2.10. *Code availability*

The code is available in our GitHub project https://github.com/Daisyli95/FlowDock. Extensive details of our experiments in the Methods and Supporting Information sections are presented to assist verification of our research.

3. Results

3.1. *FlowDock achieves superior performance on binding pose prediction*

To benchmark the performance of FlowDock against widely used docking tools, we initially trained the model on the PDBbind v2020 dataset. Subsequently, we evaluated it on the CASF 2016 dataset and the widely recognized PoseBusters test set. Empirical cumulative distribution function plots (Fig. 2A and B) depict the fraction falling below each unique RMSD value for the docked complexes. FlowDock outperforms all the baseline methods, including KarmaDock[15], DiffDock[16], Glide[9], LeDock[13], and AutoDock Vina[12] in the two benchmarks. Specifically, the method achieves a success rate of 89% for ligand RMSD below 2 Å on the CASF 2016 benchmark and 67% on the PoseBusters dataset, respectively. Fig. 2C further illustrates the overall distribution of ligand RMSD values predicted by each tool, confirming that FlowDock yields high accuracy and consistency with a higher density of predictions at lower RMSD values compared to other methods.

To take advantage of the model's capability to generate diverse conformations, we introduced the pLDDT scoring module, based on AlphaFold LDDT[50] score, aimed at selecting the optimal complex structure from the generated results. As illustrated in Fig. 2D, our predicted ligand pLDDT score has a strong correlation with native ligand RMSD, proving the efficacy in assessing complex structures. The ROC_AUC[54] using a ligand RMSD below 2 Å as the criterion for a positive sample is 0.81. Furthermore, the success rate improves as the number of samples increases, highlighting the model's ability to enhance prediction quality with increased sampling (Fig. 2E). Ligand RMSD alone is not a sufficient standard for evaluating docking quality; the physical plausibility of the generated conformation, as discussed in the Method Section, is equally critical. Our approach significantly outperforms both the leading energy-based method, Glide, and the diffusion-based method, DiffDock[16]. Notably, the fixed number of samples generated by Glide is primarily due to the elimination of physically implausible conformations. A flat line denotes the best docking result of Glide[9], reflecting the success rate of its optimal sample. FlowDock's superior performance is attributed to its capacity to accommodate substantial protein conformational changes, resulting in a more accurate fit between the protein and ligand.

FlowDock has shown the capability to reduce protein RMSD compared to the initial AlphaFold2 predicted structure (Fig. 2F). This demonstrates the method's ability to handle conformational

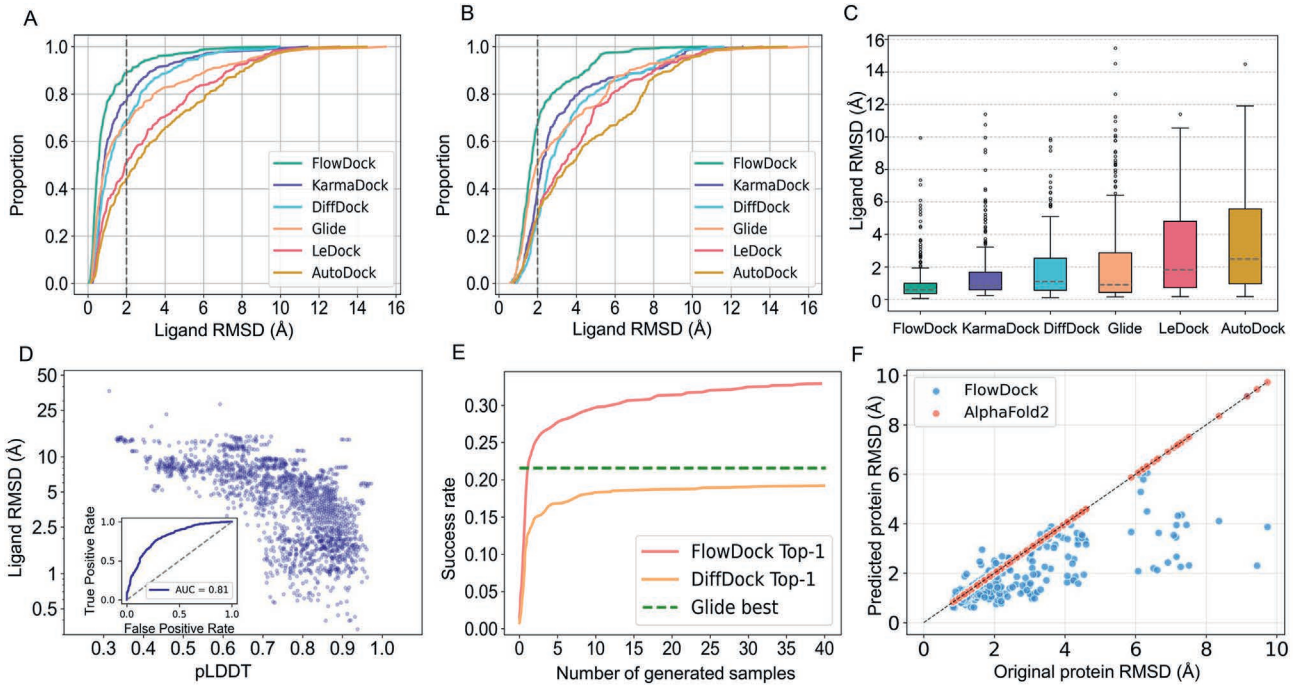


Figure 2 Performance of binding pose prediction on CASF 2016 (n= 285) and PoseBusters v2 (n= 428). (A, B) Empirical cumulative distribution function (ECDF) plots depicting the proportion of complexes with RMSD below thresholds for various tools using the CASF 2016 (A) and PoseBusters dataset (B). (C) Distribution of ligand RMSD values across different prediction tools: FlowDock, KarmaDock, DiffDock, Glide, LeDock and AutoDock. (D) Correlation between ligand RMSD and model-predicted pLDDT scores, highlighting pLDDT as an effective predictor of true ligand RMSD (ROC_AUC = 0.81). (E) Success rate (ligand RMSD <2 Å, clash rate<0.4) increases as the number of generated samples increases. (F) FlowDock predicted protein RMSD is less than the original Alphafold2 predicted protein RMSD, indicating better alignment with native crystallized structures.

changes and recover holo-structures where other approaches might struggle. While our method outperforms other approaches on both CASF 2016 and PoseBusters benchmarks, we observe that its performance is notably stronger on CASF 2016 compared to PoseBusters. This highlights potential issues with overfitting or limited generalization ability to unseen data, underscoring the need for further refinements to improve the framework's robustness. Beyond the reduced protein RMSD, we have also calculated the predicted system energy in comparison with the crystallized structure, and the results show that our system is very close to this, as illustrated in Supporting Information Fig. S1.

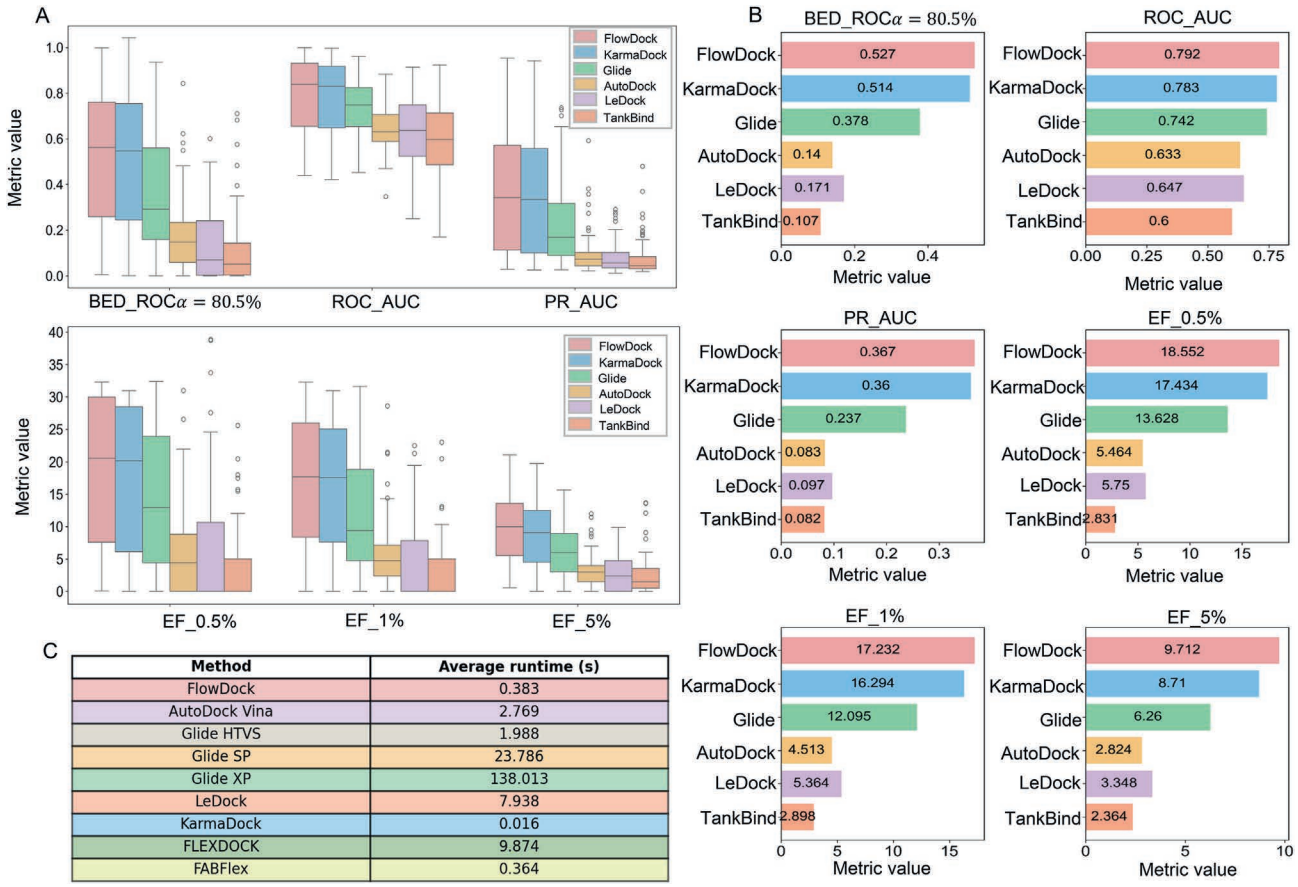
In addition, we investigated the influence of ligand complexity on binding pose accuracy. As shown in Supporting Information Fig. S2, FlowDock maintains consistent performance regardless of the number of heavy atoms or rotatable bonds in the ligand, indicating strong generalization capabilities across a broad range of molecular sizes and flexibilities. However, for exceptionally complex ligands—specifically those with more than 40 heavy atoms—we observed significant deviations in the predicted structures (Supporting Information Fig. S3). These cases were further analyzed to understand better the limitations of the model under extreme molecular conditions.

Overall, we conclude that FlowDock shows promising potential as a robust framework for protein−ligand complex refinement. Its ability to recover accurate holo-structures, generalize across diverse ligand chemistries, and outperform existing methods on standard benchmarks underscores its utility. Nevertheless, addressing outlier cases with highly complex ligands remains an area for future development, potentially through the incorporation of ligand-specific attention mechanisms or enhanced sampling strategies.

3.2. FlowDock exhibits better screening ability in terms of accuracy and speed

FlowDock's virtual screening performance was comprehensively evaluated on the DEKOIS 2.0 dataset, benchmarking it against a suite of traditional docking tools and state-of-the-art deep learning-based methods KarmaDock[15] and TankBind[43]. The results, summarized in Fig. 3, indicate that FlowDock consistently demonstrates superior screening power across all six evaluation metrics. Here, we used a new baseline, TankBind, as some of our baseline methods, DiffDock[16], FLEXDOCK[59], and FABFlex[60], were not designed for screening. Notably, our method is among the few deep learning approaches capable of both predicting binding poses and screening.

Moreover, the docking speed of each method is compared. The results demonstrate that FlowDock is significantly faster than traditional docking tools and comparable to regressive methods, KarmaDock and FABFlex, which generate a single binding pose



Method	Average runtime (s)
FlowDock	0.383
AutoDock Vina	2.769
Glide HTVS	1.988
Glide SP	23.786
Glide XP	138.013
LeDock	7.938
KarmaDock	0.016
FLEXDOCK	9.874
FABFlex	0.364

Figure 3 Screening performance on DEKOIS 2.0 dataset containing 81 targets. (A) Distributions of key metrics including BED_ROCα= 80.5%, ROC_AUC, PR_AUC, EF_0.5%, EF_1%, EF_5% obtained from the various docking models (n= 81 data points per box). (B) The average results of these metrics obtained from these models. (C) Average runtime per binding pose. Average results from five independent runs are reported.

per molecule. Better than that, FlowDock efficiently produces multiple binding poses simultaneously, potentially aiding in the identification of optimal binding configurations or interactions.

3.3. FlowDock captures conformational changes of proteins and exhibits enhanced stability during the generation phase

Conventional docking tools and deep learning docking methods treat a protein as a rigid body and often ignore its flexibility and dynamics. This leads to a huge bias between the predicted binding pose and ground-truth complex structures[31]. Our method, in contrast, takes the protein flexibility of residue changes into account and can capture the protein conformation changes as illustrated in Fig. 4A−C. As a representative example, for PDB 4llx, the generated protein RMSD is 0.326 Å, compared to a protein RMSD of 0.433 Å for the AlphaFold2-predicted structure calculated by Pymol. Similarly, for PDB 4m0y, the predicted complex RMSD is 0.349 Å, outperforming the AlphaFold2 pocket RMSD of 0.620 Å. These results indicate that FlowDock-predicted protein structures are more native-like and effectively capture the dynamic conformational changes associated with protein−ligand docking.

To evaluate the generalization ability concerning different ligand structures, we chose the ligand with a variable number of rotatable bonds as shown in Fig. 4D−F. None of these ligands was

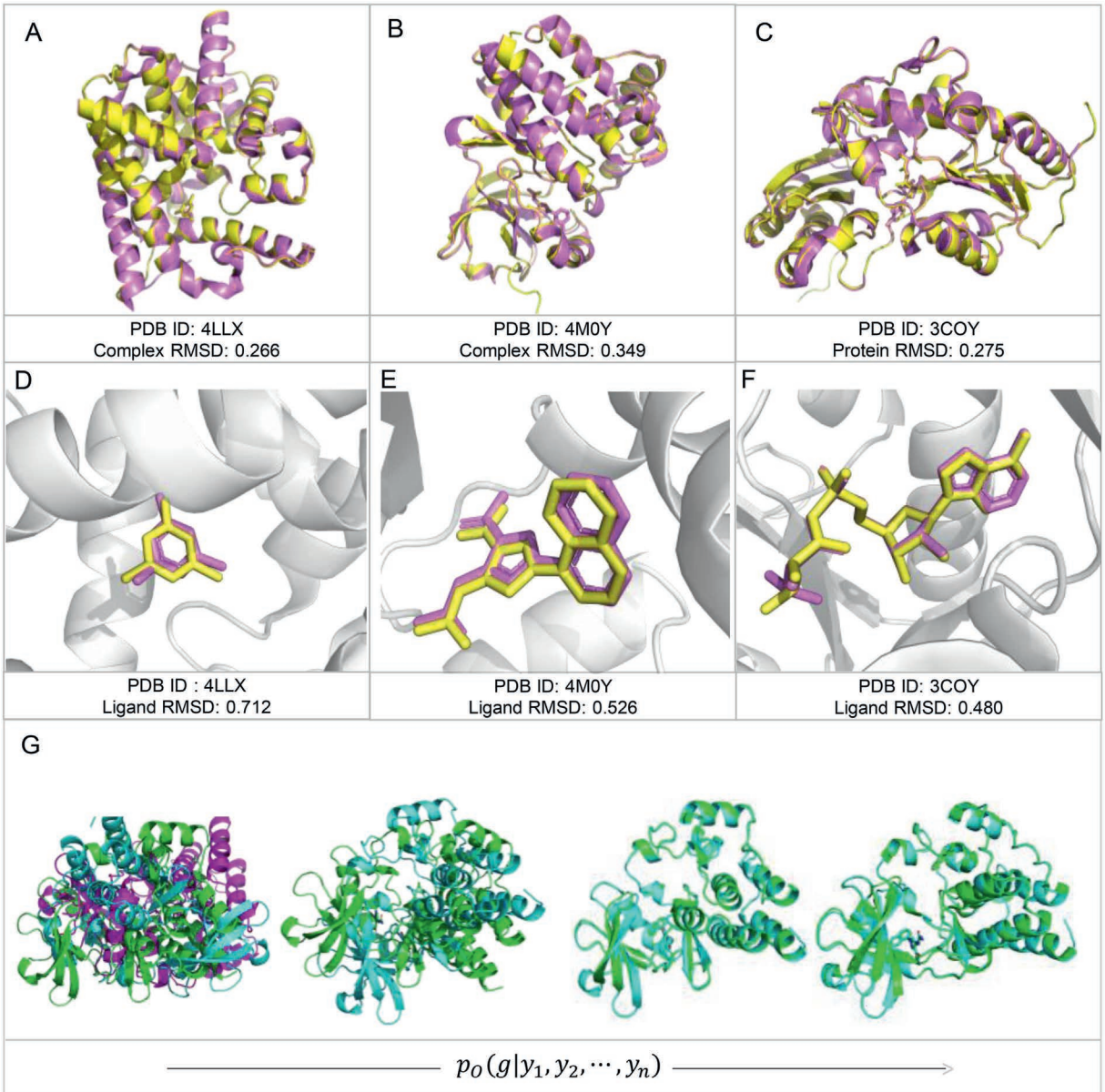


Figure 4 Illustration of FlowDock predicted results on diverse PDB and ligand structures. The ground-truth structures downloaded from PDB website are colored in yellow and the FlowDock predicted structures are colored in violet. (A−C) Predicted results of small ligands with 0, 4 and 8 rotatable bonds and 9, 22, 31 heavy atoms on PDB 4LLX, 4M0Y and 3COZ, respectively. (D−F) Local illustration on corresponding ligand conformations. All protein RMSD values, as well as ligand RMSD values, are below 2 Å. (G) Illustration of generation procedure of PDB 4M0Y with ligand. The ground-truth complex structure is colored in green, and the predicted structure is colored in cyan and magenta. Protein conformation and ligand pose keep updating as the iterative sampling continues until n steps.

used in the training set. The results show that the predicted complex structures have low ligand RMSD and low protein RMSD, regardless of the number of rotatable bonds and heavy atoms. This demonstrates the FlowDock framework's capability to capture the flexibility and dynamics of proteins and ligands, as well as its generality across diverse protein and ligand structures.

A key principle underlying FlowDock is that a model is more likely to learn and generalize patterns or structures that align well with the underlying data characteristics when the changes or transitions between intermediate steps in the generation process are more coherent and smoother[28]. The generation process transpires within the parameter space of FlowDock, whose regulation is achieved *via* the Bayesian update functions. In this procedure, samples with higher noise levels are assigned reduced weight, resulting in a marked decrease in variance within the parameter space[28]. Consequently, this enables a smoother transformation of molecular geometries, including translation, rotation, and torsion angle adjustments for both proteins and ligands. As presented in Fig. 4G, the intermediary structures progressively converge towards the final structure. This highlights the efficacy of the smoother transformation approach.

3.4. FlowDock generates more physically plausible conformations among deep learning methods

Physical plausibility is a critical criterion for evaluating the validity of predicted binding poses, yet it remains a persistent challenge for deep learning-based methods[31,47]. In this study, we assess our approach using the most widely applied benchmark dataset, PoseBusters, and demonstrate a substantial improvement in physical plausibility compared to existing deep learning models. As shown in Fig. 5, our method achieves a whole validity of over 51.2%, while other deep learning methods best achieve 40.7%. While deep learning models often struggle with physical plausibility due to limited training data and the absence of explicit physical constraints, our method significantly outperforms other state-of-the-art approaches, narrowing the gap between data-driven models and physics-based docking tools. However, there still remains a gap towards Alphafold3 with 84% whole validity. But we note that Alphafold3 was trained on a dataset 10 times larger than ours, and its computational cost is high. This evaluation is performed using the PoseBusters package, which quantitatively measures the physical realism of predicted poses. In addition to performing the standard physical validity checks on the PoseBusters dataset, we also applied the same evaluation to the recently published TrueDecoy benchmark set. This further demonstrates FlowDock's capability to generate physically plausible conformations across diverse datasets. The results are presented in Supporting Information Fig. S4.

3.5. FlowDock reaches better performance on binding affinity prediction

By effectively capturing molecular representations and intricate interactions between ligands and binding pockets, FlowDock could make accurate predictions of binding affinity. We evaluate its performance (Top-1 result) on the CASF 2016 dataset, comparing it to both complex-based and complex-free baseline methods. Complex-based methods leverage the holo-structure of protein−ligand complexes, utilizing more detailed structural information than complex-free approaches, which rely solely on ligand or protein data without the bound complex structure. The results are presented in Fig. 6. Even without relying on prior knowledge of the complex structure, like complex-based methods, FlowDock achieves comparable performance, showcasing its capability of capturing the characteristics of molecules. The comparison between FlowDock's predicted and experimental binding affinity highlights a strong alignment between the two,

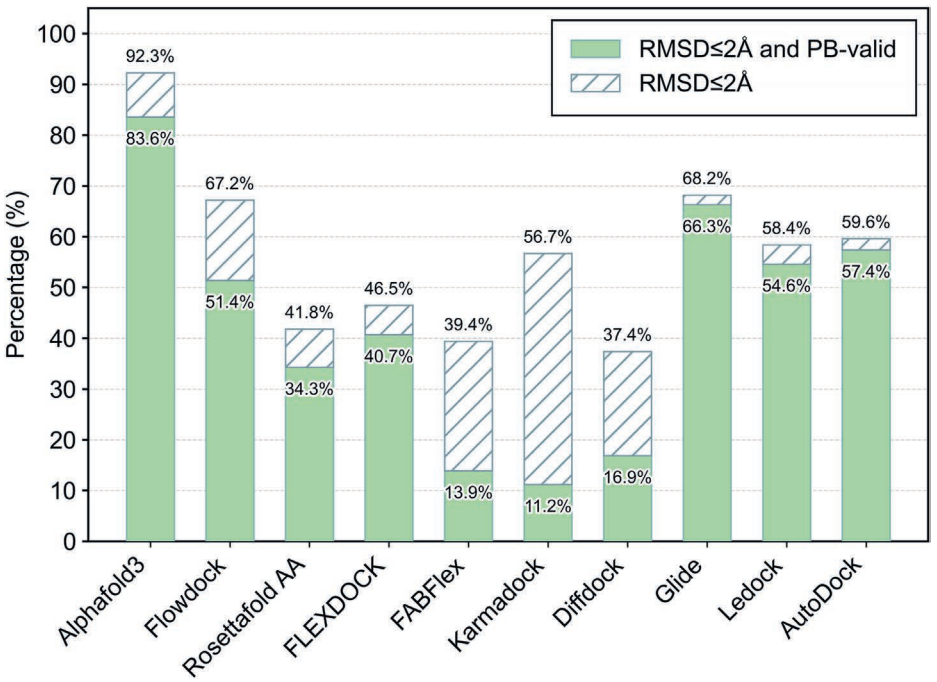


Figure 5 Performance of different docking methods on the PoseBusters v2 benchmark ($n = 428$). AlphaFold3 was trained on a dataset 10 times larger.

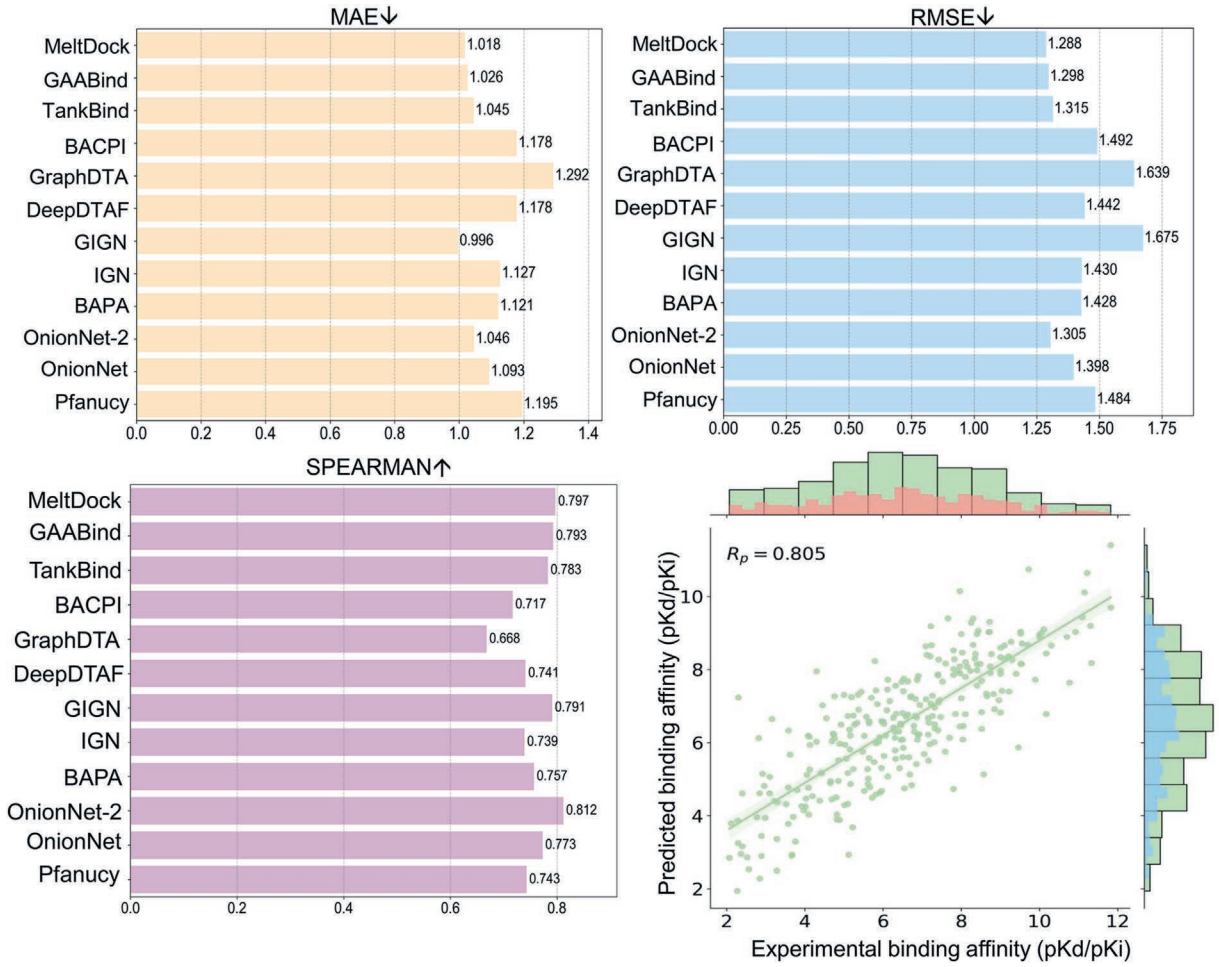


Figure 6 The performance of FlowDock on binding affinity prediction on CASF 2016 ($n = 285$). (A) MAE measures the average magnitude of the prediction error by taking the absolute difference between predicted and experimental affinities, with FlowDock's MAE being 1.018. (B) RMSE evaluates the prediction accuracy by calculating the square root of the average squared differences between predicted and experimental binding affinities. (C) Spearman's correlation coefficient assesses the rank-order correlation between predicted and experimental affinities. It evaluates how well the predictions preserve the relative ranking of binding affinities, regardless of the exact values. (A–C) Average results from five independent runs are reported. (D) Regression curve comparing predicted binding affinity with experimental binding affinity on the CASF 2016 dataset with $R^2 = 0.805$.

demonstrating a high level of correlation. As the sample size increases from 1 to 40, the ensemble average of predicted affinity exhibits a lower RMSE and shows stronger Spearman correlations with the ground truth. (Supporting Information Table S2).

3.6. Case study: application of FlowDock to HPK1 using curated experimental data

Hematopoietic Progenitor Kinase 1 (HPK1, also known as MAP4K1) is a critical negative regulator of immune cell signaling and has recently emerged as a promising target in immuno-oncology. Due to its well-characterized kinase domain and the availability of high-resolution structural data, HPK1 serves as an ideal test case for structure-based drug discovery and virtual screening approaches.

To assess the practical utility of FlowDock, we curated a dataset of over 12,000 experimentally validated HPK1 ligands from the BindingDB database, each annotated with quantitative binding affinity measurements. Ligands were classified as active or inactive using a threshold of 1 μmol/L, enabling a robust benchmarking framework for evaluating virtual screening performance. The results are shown in Fig. 7.

Firstly, we compared the HPK1 binding pocket conformation used in FlowDock with both the AlphaFold2-predicted structure and an experimentally determined crystal structure (PDB ID: 7L24). The RMSD between the FlowDock-predicted structure and the crystal structure is 0.637 Å, whereas the RMSD between the AlphaFold2-predicted structure and the crystal structure is 0.804 Å. Structural alignment revealed that FlowDock's model more accurately captured the protein's conformational state, particularly in flexible regions critical for ligand binding. This suggests that FlowDock not only enhances virtual screening performance but also benefits from a more realistic representation of protein flexibility, which is often a limiting factor in structure-based modeling.

In addition, FlowDock demonstrated strong discriminative power, as reflected in both ranking metrics and early enrichment performance, indicating its effectiveness in prioritizing true

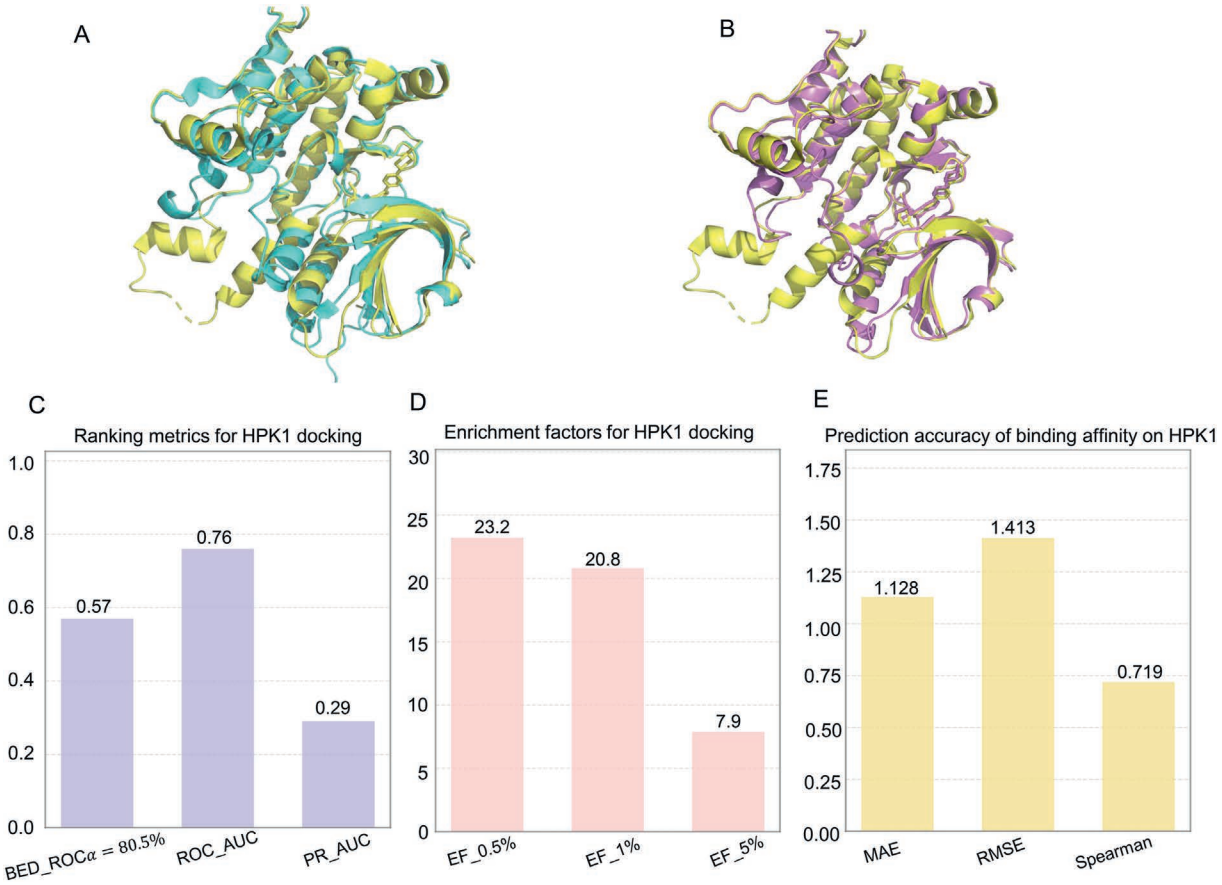


Figure 7 Benchmarking FlowDock on the HPK1 target using experimentally curated data. (A) Structural alignment between the AlphaFold2-predicted HPK1 model and the crystallographic structure (PDB ID: 7L24), highlighting deviations in the binding pocket. (B) Structural comparison between FlowDock's protein model and the same crystallographic reference, demonstrating improved conformational agreement and better modeling of flexible regions relevant for ligand binding. (C) Ranking performance of FlowDock in distinguishing active from inactive HPK1 ligands, evaluated using ROC AUC, PR AUC, and BEDROC ($\alpha = 80.5\%$). (D) Enrichment analysis showing the early retrieval performance of FlowDock at the top 0.5%, 1%, and 5% of the ranked compound library. (E) Regression analysis of predicted *versus* experimental binding affinities for HPK1 ligands. FlowDock predictions show strong quantitative accuracy, as reflected by a MAE, RMSE, and Spearman correlation. MAE, RMSE, and Spearman correlation are calculated on the full dataset ($n = 5248$) across five independent runs.

binders from an extensive compound library. These results confirm the model's ability to recover active compounds with high precision at the top ranks, a key requirement for practical drug discovery applications.

To further investigate the model's ability to predict quantitative binding affinities, we extracted a subset of 5248 compounds from this curated dataset with experimentally determined affinity values against HPK1. The corresponding log-transformed values were calculated to ensure consistency and facilitate comparison. Model performance was then evaluated using standard regression metrics, including MAE, RMSE, and Spearman's rank correlation coefficient between predicted and experimentally measured binding affinities. FlowDock achieved an MAE of 1.128, an RMSE of 1.413, and a Spearman correlation of 0.719, indicating strong agreement with experimental values. These results suggest that FlowDock not only excels in virtual screening and early enrichment but also provides reliable affinity estimation, which is essential for downstream lead optimization stages.

Together, these findings highlight FlowDock's potential as a reliable and effective tool for real-world drug discovery pipelines, particularly in the context of challenging immuno-oncology targets like HPK1.

4. Discussion

Predicting binding poses of small molecules to target proteins can be solved using two main categories of deep learning methods: regression-based approaches and probabilistic generative modeling approaches[61]. Regression-based approaches focus on optimizing metrics reflecting the probabilistic likelihood of the given data, instead of directly minimizing a loss function designed to identify the most accurate binding pose within a specific pocket. In contrast to that, generative models directly sample from complex structures and learn data distributions of ligands and proteins, often leading to higher accuracy, yet are more computationally costly when generating multiple poses. Our established framework, FlowDock, is a generative framework with a relatively low computational expense while generating rational complex structure (Ligand RMSD<2 Å, clash rate<0.4), achieving higher speed and accuracy.

In our work, we present significant advancements in addressing the multifaceted challenge of studying protein−ligand interactions. We demonstrate that FlowDock, a generative deep learning framework incorporating compound information, protein dynamics, and biophysical inductive bias[62], consistently surpasses the performance of existing molecular docking methods. Contingent upon the availability of required features, FlowDock efficiently produces high-fidelity structural predictions within seconds. This functionality enables the systematic investigation of protein−ligand interactions across extensive chemical landscapes, leveraging the continual advancements in databases of chemically diverse compound libraries and computationally predicted protein structures[57,63].

FlowDock integrates two traditionally distinct processes—complex structure generation and binding affinity prediction—into a unified framework. As a data-driven deep learning approach, FlowDock achieves computational speeds that are orders of magnitude faster than traditional docking tools, particularly in sampling extensive protein conformational changes. Without the demand of user-defined binding pockets[9,12,64], this framework is highly effective for virtual screening of compounds when binding to unknown binding sites. In this way, it significantly improves the probability of identifying compounds with selective binding affinity to the intended protein target, thereby minimizing the risk of off-target interactions—a critical consideration in rational drug design. Additionally, this methodology can aid in target deconvolution by identifying the binding partner of compounds whose biological activity was originally observed through phenotypic screening[65].

FlowDock demonstrates state-of-the-art performance across our benchmark evaluations; however, there is still scope for improvement, particularly in enhancing its ability to generalize to novel proteins not represented in the training set. More specifically, its superior performance on the CASF 2016 dataset compared to PoseBusters highlights potential overfitting and underscores the challenges of extrapolating to unseen data—limitations that are characteristic of many deep learning-based approaches. Despite these challenges, FlowDock's performance is poised to benefit substantially from the ongoing advancements in Cryo-EM techniques, which are rapidly expanding the availability and diversity of high-resolution structural data[66]. Recent developments in all-atom protein structure prediction tools, such as AlphaFold3 and RFAA, offer exciting opportunities further to enhance the modeling of protein flexibility in docking tasks. While these tools were not included in our current benchmarking due to their high computational cost, we recognize their potential to improve the accuracy of protein conformational sampling and binding site prediction. We believe the active exploration of integrating such models into future versions of FlowDock is a promising avenue. These technological advancements will enhance the diversity and comprehensiveness of our training data by providing a wider variety of protein−ligand complex conformations at an accelerated pace[58,67].

To summarize, FlowDock introduces an innovative strategy for exploring protein−ligand interactions, distinguishing itself from conventional docking techniques that consider proteins as rigid, as well as from molecular dynamics simulations, which are often computationally demanding. This method's ability to account for comprehensive protein flexibility holds substantial promise for identifying drug molecules and potential binding pockets. This multi-modal multi-task framework demonstrates its power for real-world applications to facilitate drug design.

5. Conclusions

This study presented a novel multimodal multi-task deep generative framework, FlowDock, to generate high-fidelity protein−ligand complex structures and binding affinities, with low computational cost. This newly established framework has shown a significant improvement in structural rationality, predicting binding pose. We anticipate that the tool could facilitate structure-based drug design with its high speed and accuracy.

Acknowledgments

This work is supported by the Science and Technology Development Fund, Macau, SAR, China (No. 0030/2024/RIA1) and Macao Polytechnic University (No. RP/FCA-15/2023, China). We thank Likun Zhao and Li Qin for their assistance with the use of molecular docking software.

Author contributions

Xiaojun Yao and Huanxiang Liu conceived and supervised the project, while also polishing the manuscript. Jing Li developed this methodology and wrote the manuscript. Ruiqiang Lu, Yi Tan, Pengyu Liang, and Bo Liu conducted the data analysis. Shukai Gu designed new experiments in case study during revision. All the authors discussed the experimental results, read and approved the final manuscript.

Conflicts of interest

The authors have no conflicts of interest to declare.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2026.04.009.

References

1. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 2024;**630**:493−500.
2. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* 2024;**384**. eadl2528.
3. Cao D, Chen G, Jiang J, Yu J, Zhang R, Chen M, et al. Generic protein−ligand interaction scoring by integrating physical prior knowledge and data augmentation modelling. *Nat Mach Intell* 2024;**6**:688−700.
4. Méndez-Lucio O, Ahmad M, del Rio-Chanona EA, Wegner JK. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nat Mach Intell* 2021;**3**:1033−9.
5. Min Y, Wei Y, Wang P, Wang X, Li H, Wu N, et al. From static to dynamic structures: improving binding affinity prediction with graph-based deep learning. *Adv Sci* 2024;**11**:2405404.
6. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* 2004;**3**:935−49.
7. Tran-Nguyen V-K, Junaid M, Simeon S, Ballester PJ. A practical guide to machine-learning scoring for structure-based virtual screening. *Nat Protoc* 2023;**18**:3460−511.

8. Wang L, Wang S, Yang H, Li S, Wang X, Zhou Y, et al. Conformational space profiling enhances generic molecular representation for AI-Powered ligand-based drug discovery. *Adv Sci* 2024;**11**:2403998.
9. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* 2004;**47**: 1739−49.
10. Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. *Nature* 2023;**616**:673−85.
11. Piraner DI, Abedi MH, Duran Gonzalez MJ, Chazin-Gray A, Lin A, Zhu I, et al. Engineered receptors for soluble cellular communication and disease sensing. *Nature* 2024:1−9.
12. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 2010;**31**:455−61.
13. Zhao H, Huang D, Caflisch A. Discovery of tyrosine kinase inhibitors by docking into an inactive kinase conformation generated by molecular dynamics. *ChemMedChem* 2012;**7**:1983−90.
14. Lu W, Zhang J, Huang W, Zhang Z, Jia X, Wang Z, et al. DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat Commun* 2024;**15**:1071.
15. Zhang X, Zhang O, Shen C, Qu W, Chen S, Cao H, et al. Efficient and accurate large library ligand docking with KarmaDock. *Nat Comput Sci* 2023;**3**:789−804.
16. Corso G, Stärk H, Jing B, Barzilay R, Jaakkola TS. DiffDock: diffusion steps, twists, and turns for molecular docking. In: *The eleventh international conference on learning representations*; 2022.
17. Qiao Z, Nie W, Vahdat A, Miller TF, Anandkumar A. State-specific protein−ligand complex structure prediction with a multiscale deep generative model. *Nat Mach Intell* 2024;**6**:195−208.
18. Lai H, Wang L, Qian R, Huang J, Zhou P, Ye G, et al. Interformer: an interaction-aware model for protein-ligand docking and affinity prediction. *Nat Commun* 2024;**15**:10223.
19. Xiong J, Cui R, Li Z, Zhang W, Zhang R, Fu Z, et al. Transfer learning enhanced graph neural network for aldehyde oxidase metabolism prediction and its experimental application. *Acta Pharm Sin B* 2024; **14**:623−34.
20. Wu J, Wan Y, Wu Z, Zhang S, Cao D, Hsieh C-Y, et al. MF-SuP-p*K*a: multi-fidelity modeling with subgraph pooling mechanism for p*K*a prediction. *Acta Pharm Sin B* 2023;**13**:2572−84.
21. Ackloo S, Al-awar R, Amaro RE, Arrowsmith CH, Azevedo H, Batey RA, et al. CACHE (critical assessment of computational Hit-finding experiments): a public−private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nat Rev Chem* 2022;**6**:287−95.
22. Kaplan AL, Confair DN, Kim K, Barros-Álvarez X, Rodriguiz RM, Yang Y, et al. Bespoke library docking for 5-HT2A receptor agonists with antidepressant activity. *Nature* 2022;**610**:582−91.
23. Lyu J, Wang S, Balius TE, Singh I, Levit A, Moroz YS, et al. Ultra-large library docking for discovering new chemotypes. *Nature* 2019; **566**:224−9.
24. Lewandowski JR, Halse ME, Blackledge M, Emsley L. Direct observation of hierarchical protein dynamics. *Science* 2015;**348**:578−81.
25. Ishima R, Torchia DA. Protein dynamics from NMR. *Nat Struct Mol Biol* 2000;**7**:740−3.
26. Stank A, Kokh DB, Fuller JC, Wade RC. Protein binding pocket dynamics. *Acc Chem Res* 2016;**49**:809−15.
27. Kumar A, Chauhan S. Use of simplified molecular input line entry system and molecular graph based descriptors in prediction and design of pancreatic lipase inhibitors. *Future Med Chem* 2018;**10**:1603−22.
28. Graves A, Srivastava RK, Atkinson T, Gomez F. Bayesian flow networks. *ArXiv* 2023. Available from: https://arxiv.org/abs/2308.07037.
29. Wang R, Fang X, Lu Y, Wang S. The PDBbind database: collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J Med Chem* 2004;**47**:2977−80.
30. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative assessment of scoring functions: the CASF-2016 update. *J Chem Inf Model* 2019;**59**:895−913.
31. Buttenschoen M, Morris G M, PoseBusters M Deane C. AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem Sci* 2024;**15**:3130−9.
32. Gu S, Shen C, Zhang X, Sun H, Cai H, Luo H, et al. Benchmarking AI-powered docking methods from the perspective of virtual screening. *Nat Mach Intell* 2025;**7**:509−20.
33. Bauer MR, Ibrahim TM, Vogel SM, Boeckler FM. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0 − a public library of challenging docking benchmark sets. *J Chem Inf Model* 2013;**53**:1447−62.
34. Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P. Development and evaluation of a deep learning model for protein−ligand binding affinity prediction. *Bioinformatics* 2018;**34**:3666−74.
35. Seo S, Choi J, Park S, Ahn J. Binding affinity prediction for protein−ligand complex using deep attention mechanism based on intermolecular interactions. *BMC Bioinf* 2021;**22**:542.
36. Zheng L, Fan J, Mu Y. OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein−ligand binding affinity prediction. *ACS Omega* 2019;**4**:15956−65.
37. Wang Z, Zheng L, Liu Y, Qu Y, Li Y-Q, Zhao M, et al. OnionNet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Front Chem* 2021;**9**:753002.
38. Jiang D, Hsieh C-Y, Wu Z, Kang Y, Wang J, Wang E, et al. InteractionGraphNet: a novel and efficient deep graph representation learning framework for accurate protein−ligand interaction predictions. *J Med Chem* 2021;**64**:18209−32.
39. Yang Z, Zhong W, Lv Q, Dong T, Yu-Chian Chen C. Geometric interaction graph neural network for predicting protein−ligand binding affinities from 3D structures (GIGN). *J Phys Chem Lett* 2023;**14**: 2020−33.
40. Wang K, Zhou R, Li Y, Li M. DeepDTAF: a deep learning method to predict protein−ligand binding affinity. *Briefings Bioinf* 2021;**22**. bbab072.
41. Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug−target binding affinity with graph neural networks. *Bioinformatics* 2021;**37**:1140−7.
42. Li M, Lu Z, Wu Y, Li Y. BACPI: a bi-directional attention neural network for compound−protein interaction and binding affinity prediction. *Bioinformatics* 2022;**38**:1995−2002.
43. Lu W, Wu Q, Zhang J, Rao J, Li C, Zheng S. TANKBind: trigonometry-aware neural networks for drug-protein binding structure prediction. In: *Proceedings of the 36th international conference on neural information processing systems*; 2022.
44. Tan H, Wang Z, Hu G. GAABind: a geometry-aware attention-based network for accurate protein−ligand binding pose and binding affinity prediction. *Briefings Bioinf* 2023;**25**. bbad462.
45. Zhou G, Gao Z, Ding Q, Zheng H, Xu H, Wei Z, et al. Uni-Mol: a universal 3D molecular representation learning framework. In: *The eleventh international conference on learning representations*; 2022.
46. Cheng L, Liu F, Yao D (Daphne). Enterprise data breach: causes, challenges, prevention, and future directions. *WIREs Data Min Knowl Discovery* 2017;**7**:e1211.
47. Gu S, Shen C, Zhang X, Sun H, Cai H, Luo H, et al. Benchmarking AI-powered docking methods from the perspective of virtual screening. *Nat Mach Intell* 2025;**7**:509−20.
48. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123−30.
49. RDKit Landrum G. A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* 2013;**8**: 5281.
50. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583−9.
51. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* 2022;**13**:2453.

52. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. In: *Proceedings of the 34th international conference on neural information processing systems*; 2020.
53. Song Y, Gong J, Zhou H, Zheng M, Liu J, Ma W-Y. Unified generative modeling of 3D molecules with Bayesian flow networks. In: *The twelfth international conference on learning representations*; 2023.
54. Truchon J-F, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "Early Recognition" problem. *J Chem Inf Model* 2007;**47**:488−508.
55. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, et al. Diverse, high-quality test set for the validation of protein−ligand docking performance. *J Med Chem* 2007;**50**:726−41.
56. Yin Y, Hu H, Yang Z, Xu H, Wu J. RealVS: toward enhancing the precision of top hits in ligand-based virtual screening of drug leads from large compound databases. *J Chem Inf Model* 2021;**61**: 4924−39.
57. Zhang X, Shen C, Zhang H, Kang Y, Hsieh C-Y, Hou T. Advancing ligand docking through deep learning: challenges and prospects in virtual screening. *Acc Chem Res* 2024;**57**:1500−9.
58. Bender BJ, Gahbauer S, Luttens A, Lyu J, Webb CM, Stein RM, et al. A practical guide to large-scale docking. *Nat Protoc* 2021;**16**: 4799−832.
59. Corso G, Somnath VR, Getz N, Barzilay R, Jaakkola T, Krause A. Composing unbalanced flows for flexible docking and relaxation. In: *The thirteenth international conference on learning representations*; 2024.
60. Zhang Z, Wu L, Gao K, Yao J, Qin T, Han B. Fast and accurate blind flexible docking. In: *The thirteenth international conference on learning representations*; 2024.
61. Atas Guvenilir H, Doğan T. How to approach machine learning-based prediction of drug/compound−target interactions. *J Cheminf* 2023;**15**:16.
62. AlQuraishi M, Sorger PK. Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat Methods* 2021;**18**:1169−80.
63. Agamah FE, Mazandu GK, Hassan R, Bope CD, Thomford NE, Ghansah A, et al. Computational/*in silico* methods in drug target and lead prediction. *Briefings Bioinf* 2020;**21**:1663−75.
64. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein−ligand docking using GOLD. *Proteins: Struct, Funct, Bioinf* 2003;**52**:609−23.
65. Moffat JG, Rudolph J, Bailey D. Phenotypic screening in cancer drug discovery — past, present and future. *Nat Rev Drug Discov* 2014;**13**: 588−602.
66. Yip KM, Fischer N, Paknia E, Chari A, Stark H. Atomic-resolution protein structure determination by cryo-EM. *Nature* 2020;**587**: 157−61.
67. Nandy A, Duan C, Taylor MG, Liu F, Steeves AH, Kulik HJ. Computational discovery of transition-metal complexes: from high-throughput screening to machine learning. *Chem Rev* 2021;**121**: 9927−10000.