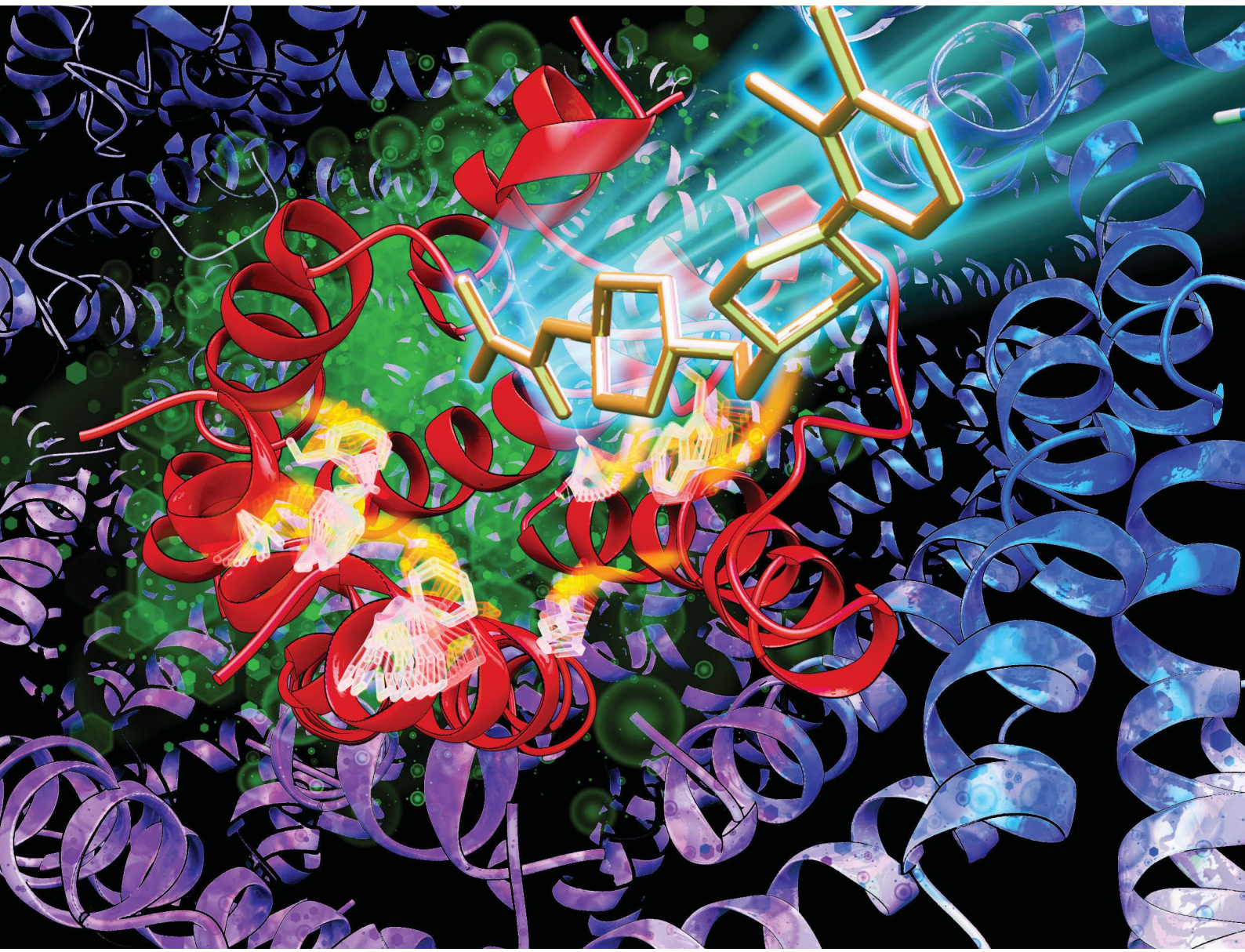
Volume 15
Number 21
7 June 2024
Pages 7797–8252

# Chemical Science

rsc.li/chemical-science



ISSN 2041-6539

ROYAL SOCIETY OF CHEMISTRY

**EDGE ARTICLE**

Jintao Zhu, Zhonghui Gu, Jianfeng Pei and Luhua Lai
DiffBindFR: an SE(3) equivariant network for flexible protein–ligand docking

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE



Cite this: *Chem. Sci.*, 2024, **15**, 7926

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 18th December 2023
Accepted 7th April 2024

DOI: 10.1039/d3sc06803j

rsc.li/chemical-science

# DiffBindFR: an SE(3) equivariant network for flexible protein–ligand docking†

Jintao Zhu, ‡[a] Zhonghui Gu, ‡[b] Jianfeng Pei *[a] and Luhua Lai *[abcd]

Molecular docking, a key technique in structure-based drug design, plays pivotal roles in protein–ligand interaction modeling, hit identification and optimization, in which accurate prediction of protein–ligand binding mode is essential. Conventional docking approaches perform well in redocking tasks with known protein binding pocket conformation in the complex state. However, in real-world docking scenario without knowing the protein binding conformation for a new ligand, accurately modeling the binding complex structure remains challenging as flexible docking is computationally expensive and inaccurate. Typical deep learning-based docking methods do not explicitly consider protein side chain conformations and fail to ensure the physical plausibility and detailed atomic interactions. In this study, we present DiffBindFR, a full-atom diffusion-based flexible docking model that operates over the product space of ligand overall movements and flexibility and pocket side chain torsion changes. We show that DiffBindFR has higher accuracy in producing native-like binding structures with physically plausible and detailed interactions than available docking methods. Furthermore, in the Apo and AlphaFold2 modeled structures, DiffBindFR demonstrates superior advantages in accurate ligand binding pose and protein binding conformation prediction, making it suitable for Apo and AlphaFold2 structure-based drug design. DiffBindFR provides a powerful flexible docking tool for modeling accurate protein–ligand binding structures.

## 1 Introduction

The primary paradigm of drug discovery involves identifying and designing molecules that target key proteins within disease pathways. Historically, screening compound libraries using biochemical platforms has been the predominant approach for identifying novel drugs.[1] Since the 1990s, high-throughput screening (HTS) has been employed on libraries ranging from 500 000 to $10^8$ molecules,[2,3] leading to the discovery of several drugs. While the HTS libraries represent a significant advancement over traditional lab-designed ones, they encompass only a fraction of potential drug-like molecules.[4] Given the challenges and expenses associated with synthesizing such a vast chemical space, computational methods for screening virtual libraries are frequently employed in drug discovery, allowing exploration of chemical spaces comprising tens of billions of molecules or even more.[5–7]

Structure-based virtual screening (SBVS) enables rapid and cost-effective modeling of target-molecule binding structures from large-scale compound libraries together with the evaluation of their binding affinities for identifying potential hits.[8–10] Molecular docking is one of the most frequently employed techniques for SBVS, which is utilized to predict ligand binding poses, characterize protein–ligand binding strength, and identify key interactions.[11,12] In general, conventional docking approaches, including AutoDock4,[13] AutoDock Vina,[14,15] Smina,[16] Glide,[17] and GOLD,[18] leverage heuristic search algorithms, to explore a variety of potential ligand conformations. Scoring functions with simplified terms are utilized for fast estimation of binding affinity and priority of ligand poses.

Classical molecular docking methods describe protein–ligand interactions based on the lock-and-key model,[19] wherein a rigid receptor binding pocket serves as the "lock" and the molecular docking algorithm primarily optimizes the ligand's conformation to find a complementary "key". Such rigid receptor docking methods, for the trade-off between accuracy and computational efficiency, strive to determine the optimal and complementary binding conformation. When known complex structures are available, and ligand molecules are removed and then re-docked into the native Holo pockets, rigid

[a]*Center for Quantitative Biology, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China. E-mail: lhlai@pku.edu.cn; jfpei@pku.edu.cn*

[b]*Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, 100871, China*

[c]*BNLMS, College of Chemistry and Molecular Engineering, Peking University, Beijing, 100871, China*

[d]*Peking University Chengdu Academy for Advanced Interdisciplinary Biotechnologies, Chengdu, Sichuan, China*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d3sc06803j

‡ These authors contributed equally to this work.

receptor docking often achieves impressive success rate.[20] However, in real-world docking tasks without knowing the binding conformation in advance, ligand induced pocket conformational changes may produce wrong docking results.[21] Virtual screening and rational design on the unbound (Apo) or computationally modeled structures usually give unsatisfied hit rate.[22–27] Without accounting for pocket flexibility, the performance of docking methods experiences drastic decrease in such cases,[28] which may rule out potential hits during the early stage of drug discovery. Although AlphaFold[29] is capable of accurately modeling target protein structures, traditional docking methods that overlook potential side chain flexibility perform less effectively when applied to the predicted structures.[30]

Currently, there are two main strategies to address the flexibility of protein pockets. The first approach involves inducing local conformational rearrangements of the target using force field or scoring function-based calculations. For instance, rDock[31] allows movements of functional groups that can form hydrogen bonds including –OH and $–NH_3$. AutoDockFR (AutoDock for Flexible Receptors)[32] allows users to specify up to 14 flexible side chains in advance and samples reasonable side chain dihedral angles from a rotamer library. Despite its better performance than AutoDock Vina in cross-docking benchmarks with Apo structures, it is considerably time-consuming and requires prior knowledge of potentially critical side chains in the pocket, which limits its application in SBVS. These methods consider partial flexibility of pockets and are suitable to handle the cases with only minor conformational rearrangements mainly resulting from side chain movements. There are a few methodologies that have been carefully devised to account for full flexibility, primarily categorized into ensemble docking and induced fit docking.[33] Ensemble docking aims to implicitly mimic the dynamical behavior of a receptor's binding site represented by a conformational ensemble.[34] The most common implementation is straightforward by docking each ligand to multiple rigid conformations of a receptor. The conformational ensemble could be obtained either by crystallographic structures or structure models derived through computational methods such as molecular dynamics (MD) simulation[35] and normal modes.[36] Clearly, ensemble docking depends significantly on the diversity and structural quality of the conformational ensemble for well-defined representation of the receptor's flexibility. It has been observed to be ineffective in identifying the true receptor conformations derived from clustering.[37] Instead, Schrödinger IFD-MD,[38] one of the rigorous induced fit docking workflows, explicitly emulates the interactions between receptors and ligands. IFD-MD elaborately integrates pharmacophore docking, rigid receptor docking, energy-driven binding site refinement, and MD simulations for multiple iterations.[38] IFD-MD requires a template pose which can be obtained either by docking a known binder to the target, or by aligning the structure to a protein homologue and grafting the post-aligned ligand coordinates. Such a knowledge-based approach adroitly leverages the available protein–ligand interaction patterns or homologous ones, which ensures the generation of a reliable initial pose. The placement often results in clashes between the ligand and the pocket, which can be resolved later in the side chain repacking and backbone minimization. Subsequently, multiple independent MD simulations and metadynamics simulation for the assessment of pose stability are run. Finally, the final poses are ranked by a composite scoring function. It is evident that this workflow burdens a significant allocation of computational resources.[38,39] Despite the efforts to relax the systems, it generally fails in cases involving backbone motion.[38,39]

The second approach is the recently developed deep learning-based methods,[40] which coarsen the representation of protein pockets by only encoding the protein backbone atoms without explicitly including side chain atoms. This kind of representation is insensitive to minor pocket backbone flexibility and side chain adaptability. Earlier works, like DeepDock,[41] TankBind,[42] and EDM-Dock,[43] predicted pocket residue-ligand distance map, which is used to reconstruct the binding structure. Leveraging powerful equivariant neural networks like EGNN,[44] geometric deep learning[45] models such as EquiBind,[46] LigPose,[47] E3Bind,[48] Uni-Mol,[49] and KarmaDock[50] iteratively predict the three-dimensional coordinates of ligands directly around the whole protein (blind docking) or predefined pocket. Recent SOTA blind docking method DiffDock,[51] based on the diffusion generative modelling,[52] employed the SE(3) equivariant neural network[53] to denoise the rotation, translation, and bond torsion of ligand, and then rank poses by additional confidence model. However, these existing deep learning-based docking approaches face limitations in effectively handling protein flexibility and the generated ligand poses are often implausible.[54] The generated ligand structures often contain clashes with the target and irrational bond lengths, angles, and torsion angles that lead to high intra energies. Ligand conformational optimization using tools such as RDKit alignment[50,55] cannot completely alleviate ligand and protein clashes. Furthermore, ignoring the target flexibility and validity of ligand poses makes it challenging for these deep learning-based methods to capture key interactions in docking.[54] Recently, building upon the methodology of optimizing ligand coordinate recycling as developed in LigPose and KarmaDock, a deep learning-based flexible docking method named FlexPose[56] has extended its predictive capabilities of pocket side chain coordinates. This advancement allows for more details of interaction information in cross-docking applications. However, like LigPose and KarmaDock, FlexPose encounters the same inherent limitation. The methodological focus on fitting coordinates within Euclidean space tends to overfit the overall RMSD (Root Mean Square Deviation). Consequently, FlexPose, akin to its predecessors, is inevitably limited by conformational rationality. Overall, these inadequacies of current methods impede subsequent steps, such as post-optimization of ligands by experts based on the detailed interactions or conducting further studies through molecular dynamics simulations.

Early Apo–Holo pair analysis has shown general consensus that upon ligand binding protein pocket undergoes significant side chain conformation heterogeneity while backbone is relatively rigid in most cases.[57–59] Therefore, in most cases, side chain flexibility modelling is enough for flexible docking. In this study, we developed a full-atom flexible docking model,

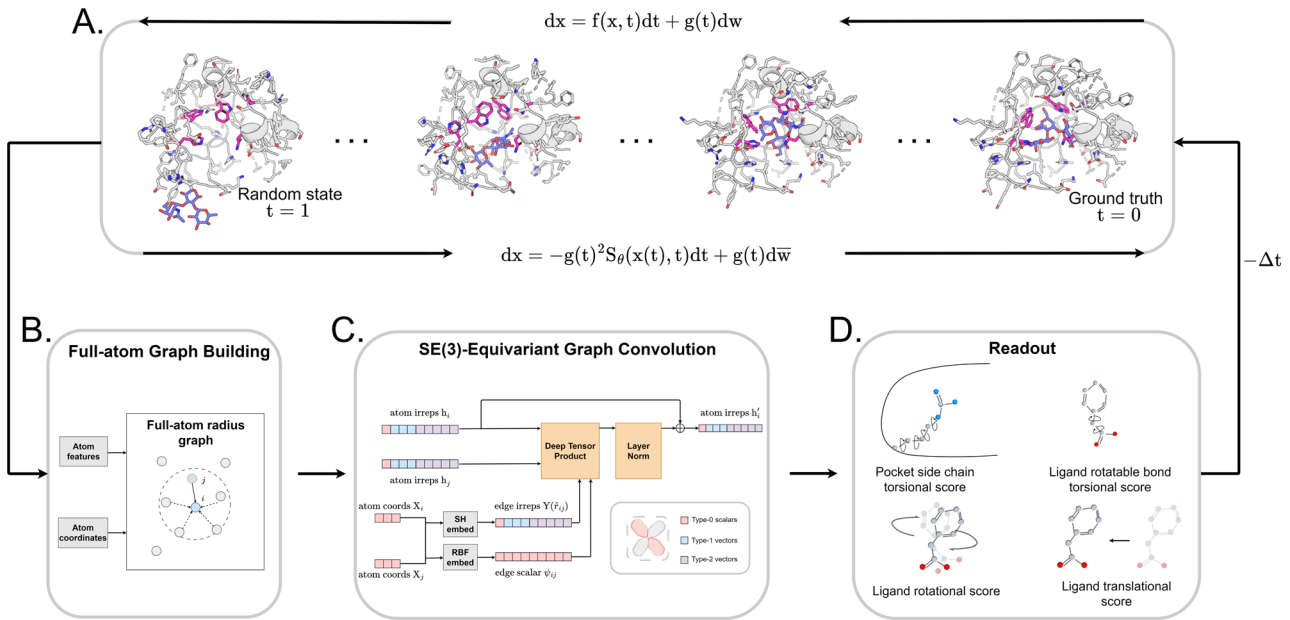



Fig. 1 The architecture of DiffBindFR. (A) Overview of score-based generative modeling through SDE for flexible docking. The flexible docking process is decomposed into ligand translation, rotation, bond torsion and pocket side chain torsion. (B) Construction of full-atom interaction graph. According to the real-time coordinate of each atom, we build the spatial graph as model input. (C) The architecture of SE(3) equivariant graph convolution. It serves as the trunk block of DiffBindFR network. $h_i$ and $X_i$ are the irreducible representations (Irreps) and coordinate of atom i, respectively. The distance and vector between atom i and atom j are embedded through Gaussian radial basis (RBF) and spherical harmonics basis (SH) respectively, to get their edge scalar representations $\psi_{ij}$ and edge vector Irreps $Y(\hat{r}_{ij})$. Then, Deep Tensor Product from e3nn library is served as message-passing module to gather messages from neighborhood, followed by an equivariant Layer Normalization (Layer Norm) module to get the updated Irreps $h_i'$. (D) The output readout of DiffBindFR network contains the predicted score of pocket side chain torsion, ligand rotatable bond torsion, ligand rotation and ligand translation. These scores are used to solve the reverse SDE for binding structures sampling.

DiffBindFR, based on the diffusion framework (Fig. 1). In the comprehensive evaluation, starting from pocket conformations with randomized side chain torsion angles, DiffBindFR outperforms state-of-the-art (SOTA) deep learning techniques and traditional docking methods. Owing to the explicit full-atom modeling of pocket residues, and its learning of joint optimization of variables within the entire system across a product space composed of torsional angles, rotations, and translations, DiffBindFR can not only accurately recover protein pocket side chain conformations, but also generate precise and highly physically plausible ligand binding poses. In cross-docking benchmark, DiffBindFR, due to its capability of full pocket side chain optimization, significantly outperforms traditional flexible docking methods like IFD-MD,[38] AutoDock VinaFlex[32] and rDock.[31] It also gives superior performance in generating both accurate and valid binding structures compared to existing docking methods in the Apo dataset and AlphaFold2 modeled structures.

## 2 Overview of DiffBindFR

We formulated flexible docking as a problem of learning the joint denoising process of four variables in their tangent space: ligand rotation R, translation T, rotatable bond torsion $\tau$, and pocket side chain torsion $\chi$. Following the VE-SDE (variance exploding stochastic differential equation) paradigm,[60] starting from the crystal complex $P(x(0)) = P(R(0), T(0), \tau(0), \chi(0))$, the forward process of the diffusion model, $P(x(t)|x(0))$, involves uniformly and continuously sampling time step $t \in [0, 1]$ and injecting noise to the four kinds of movement operator to achieve binding structure perturbation. DiffBindFR is an SE(3)-equivariant generative model, following the message-passing paradigm[61] of graph neural network, that encodes the intricate interactions between the full-atom pocket and ligand, and predicts the scores $\nabla_{x(t)} \log P_t(x(t))$.[60] In the docking procedure, starting from the randomly initialized binding conformation, the scores predicted by DiffBindFR are used to solve the reverse VE-SDE process[60] to implement denoising sampling. With physics-based scoring function Smina[16] or mixture density neural network (MDN)[41] serving as confidence model, binding structures sampled by DiffBindFR can be ranked, and then the top-1 complex pose can be selected as the final prediction.

### 2.1 Diffusion generative model

The diffusion model utilize the framework of stochastic differential equations[62] to diffuse the data distribution described as follows:

$$\mathrm{d}x = f(x,t)\mathrm{d}t + g(t)\mathrm{d}w \quad (1)$$

For $x \in \mathbb{R}^D$, $f(x,t) \in \mathbb{R}^{D\times D}$ denotes a vector-valued function called the drift coefficient of $x(t)$, and $g(t) \in \mathbb{R}^{R\times R}$ denotes a scalar

function called the diffusion coefficient of $x(t)$. The lack of canonical local coordinate system defined for ligand molecules, makes the drift coefficient hard to design for the ligand rotation. Consequently, the drift coefficient $f(x,t)$ is set to be 0, and the model becomes the score-based generative model.[60] The reverse diffusion running backwards in time, which is also known as the denoising process, is given by the following reverse-time SDE:

$$\mathrm{d}x = -g(t)g(t)^{\mathrm{T}}\nabla_{x(t)}\log P_t(x)\mathrm{d}t + g(t)\mathrm{d}\bar{w} \quad (2)$$

To estimate $\nabla_{x(t)}\log P_t(x)$, we can train a score-based neural network $S_\theta(x(t),t)$ to fit it. The standard score-match loss function is as follows:

$$J(x) = \mathbb{E}_t\Big[\lambda(t)\mathbb{E}_{x(t)\sim P_{t|0}(x(t)|x(0))} \Big[\|S_\theta(x(t),t) - \nabla_{x(t)}\log P_{t|0}(x(t)|x(0))\|^2\Big]\Big] \quad (3)$$

$\lambda(t) = 1/\mathbb{E}_{x(t)\sim P_{t|0}(x(t)|x(0))}[\|\nabla_{x(t)}\log P_{t|0}(x(t)|x(0))\|^2]$ is the pre-computed weight factors.

### 2.2 Pose transformations and diffusion on the product space

We choose the specific SDE for forward diffusion process as follows:

$$\mathrm{d}x = \sqrt{\frac{\mathrm{d}\sigma^2(t)}{\mathrm{d}t}}\,\mathrm{d}w, \quad \text{where } \sigma(t) = {\sigma_{\min}}^{1-t}{\sigma_{\max}}^{t}, \ t\in[0,\ 1] \quad (4)$$

$\sigma(t) = \{\sigma_{\mathrm{R}}(t), \sigma_{\mathrm{T}}(t), \sigma_\tau(t), \sigma_\chi(t)\}$ denotes the noises that injected into ligand rotation R, translation T, rotatable bond torsion $\tau$ and pocket side chain torsion $\chi$. According to the specific group that each variable lies in, we would design the form of corresponding $\sigma$ carefully for diffusion kernel and the score computation. For a ligand pose with $n$ atoms, $X_{\mathrm{l}}\in\mathbb{R}^{3\times n}$, translation of a ligand pose $\mathrm{T}\in\mathbb{R}^3$ lies in the 3D translation group $\mathbb{T}(3)$. The diffusion kernel for ligand is a Gaussian function with variance $\sigma_{\mathrm{T}}$ as follows, which is also utilized for computing the score of ligand translation $\nabla P_{\mathrm{t|0}}(X_{\mathrm{l}}(\mathrm{t})|X_{\mathrm{l}}(0))$:

$$P_{t|0}(X_{\mathrm{l}}(t)|X_{\mathrm{l}}(0)) = \mathcal{N}(X_{\mathrm{l}}(0), \sigma_{\mathrm{T}}(t)) \quad (5)$$

As rotation of a ligand pose lies in the 3D rotation group $\mathbb{SO}(3)$, $IGSO(3)$ distribution[63,64] was chosen as the diffusion kernel. In specific, rotation matrix $\mathrm{R}\in\mathbb{R}^{3\times3}$ can be split into a unit vector $\hat{\omega}\in\mathfrak{so}(3)$ uniformly as the rotation axis and a axis-angle $\omega\in[0,\pi]$. Consequently, the functionality of $\sigma_{\mathrm{R}}$ can be replaced by $\hat{\omega}$ and $\omega$. The diffusion kernel for ligand rotation is as follows:

$$P_{t|0}(R(t)|R(0)) = \mathbf{R}(\hat{\omega},\omega)R(0) \quad (6)$$

The score of rotation diffusion can be computed according to

$$\nabla\ln P_t(R(t)|R(0)) = \left(\frac{\mathrm{d}}{\mathrm{d}\omega}\log f(\omega)\right)\hat{\omega} \quad (7)$$

$$f(\omega) = \sum_{l=0}^{\infty}(2l+1)\exp\big(-l(l+1)\varepsilon^2/2\big)\frac{\sin((l+1/2)\omega)}{\sin(\omega/2)} \quad (8)$$

where $\varepsilon$ is a scalar variance for parameterizing the $IGSO(3)$ distribution. Torsion of pocket side chains and ligand rotatable bonds lie in the $SO(2)^m$ group and $SO(2)^k$ group respectively, where $m$ and $k$ denote the number of all $\chi$ from the pocket side chain and all $\tau$ from the ligand. Since each torsion angle coordinate lies in $[0, 2\pi)$, the $m$ torsion angles of a conformer define a hypertorus $\mathbb{T}^m$. We introduced the diffusion kernel from the work of Torsional Diffusion[65] to satisfy angle periodicity, and compute its score $\nabla P_{t|0}(\chi(t)|\chi(0))$ as follows:

$$P_{t|0}(\chi(t)|\chi(0)) \propto \sum_{d\in\mathbb{Z}^m}\exp\left(-\frac{\|\chi(0)-\chi(t)+2\pi d\|^2}{2{\sigma_\chi}^2(t)}\right) \quad (9)$$

Torsion of ligand rotatable bonds are dealt with the same way as pocket side chains.

Following the eqn (3), the loss function is set as follows:

$$J(x) = J(R) + J(T) + \sum_1^k J(\tau) + \sum_1^m J(\chi) \quad (10)$$

The forward diffusion and reverse diffusion are both performed in the product space[66] of $\mathbb{T}(3)\times SO(3)\times SO(2)^k\times SO(2)^m$, corresponding to the aforementioned four kinds of transformation.

During the forward diffusion process, we would sample $t\in[0,1]$ for each pocket–ligand pair, and then utilize the defined diffusion kernel to sample each transformation. The torsions of ligand and pocket side chains are first applied to the pose, followed by translation and rotation.

The starting point of the denoising stage is a ligand conformation generated by RDKit[55] and pocket side chains, with each type of transformation sampling from their $\sigma_{\max}$. According to eqn (2), we update complex pose using the predicted score for each type of transformation. After applying the translation and rotation matrix constructed from predicted score, torsion angles get updated. It is noteworthy that there exists entanglement between ligand translation/rotation and its bond torsion, ligand pose need to be re-aligned after bonds are twisted, which will lead to model-unaware structural alignment error. With the sampled pocket side chains fixed, we perform fast local energy relaxation on the pose using Smina[16] for error correction, obtaining the final binding conformations. The number of the denoising steps is defined as 22, and 40 poses are sampled for each pocket–ligand pair, which takes in average 50 s when the batch size is set to 16 on a single 32 GB NVIDIA Tesla V100-SXM2 GPU card.

### 2.3 Confidence model

We have explored two approaches to rank the poses generated by DiffBindFR. First, the traditional scoring function Smina is utilized to quickly score the generated full-atom pocket–ligand poses. Second, a deep learning-based scoring model based on mixture density network (MDN) is trained to fit the distance distribution between ligand atoms and pocket residues. The architecture of our MDN model is similar to the scoring module of KarmaDock, and it shares the similar input representations with DiffBindFR. To better cater for the full-atom complex

system, we set the distance pairs as each ligand atom with their nearest atoms from each pocket residues. More details of MDN model can be found in ESI Section 5.†

# 3 Datasets and evaluation metrics

## 3.1 Datasets

**3.1.1 PDBbind time-split dataset.** We use the PDBbind v2020 dataset[67] for training and evaluation. For each target protein–ligand pair within the PDBbind v2020 dataset, we define the protein pocket as any residues within 12 Å buffer of any heavy atoms in the ligand molecule. Following the time split strategy proposed by the work of EquiBind,[46] where 363 complex structures uploaded later than 2019 serve as test set. After removing ligands that exist in the test set, the remaining 16 739 structures are used for training and 968 structures are used for validation. The dataset is named as "PDBbind time-split dataset" in the article.

The time-split of PDBbind is supported to be more strict and reasonable with the protein sequence similarity of 0.484 between test set and training&validation set, compared to CASF2016-split whose protein sequence similarity is 1.00 (ESI Table S3†). Volkov *et al.*[68] have shown that the time-split of PDBbind is more practical and critical over artificial splits such as ligand scaffolds or protein sequence/structure similarity for the model generalization in drug repurposing, lead optimization, and virtual screening.

**3.1.2 Posebusters test set.** The PoseBusters test set[54] is a meticulously curated collection of crystal complexes sourced from the PDB.[69] This set encompasses a diverse array of high-caliber, recent protein–ligand complexes characterized by drug-like molecules. With 428 distinct complexes, inclusive of unique proteins and ligands released since 2021, it ensures no overlap with the complexes found in the PDBbind v2020 dataset.

**3.1.3 CD test set.** Given the current absence of a large-scale benchmark dataset for cross-docking, especially to address various cross-docking scenarios (including Apo–Holo and cross-docking between different Holo states), we have established a benchmark dataset tailored for the cross-docking evaluation, termed CD test set. We integrated ApoRef,[24] a test set constructed by constrained MD for inducing Apo-like pockets into Holo-like pockets; several prominent ensemble docking targets including CDK2, EGFR, FXA; CASF2016;[67] DUDE27-HoloEns consisting of Holo structure ensembles from 27 targets in DUD-E[70] filtered by Zhang *et al.*[26] and GPCR-AF2 [30] that contains 18 human GPCR complexes published after April 30, 2018. ApoBind,[71] AHoJ[72] and SIENA[73] are utilized to search for corresponding Apo and Holo states based on queried Holo structures, thereby creating pairs for the Apo–Holo and Holo–Holo mixed cross-dock dataset. The detailed protocol for constructing the CD test set can be found in ESI Section 1.† The finalized CD test set comprises of 14 462 structural pairs for cross-docking benchmark tests.

**3.1.4 DUDE27-AF2 test set.** The DUDE27-AF2 test set is an induced fit docking test set from the work of Zhang *et al.*,[26] which contains the Holo and AF2 predicted structures of 27 targets from DUD-E[70] that are suitable for IFD-MD refinement.[38] It also contains refined AF2 modeled structures using the IFD-MD based on the ligand template either by the ideal aligned Holo ligand or docked pose sampled in the Apo pocket using Glide. See details in ESI Table S10.†

## 3.2 Evaluation metrics

We utilize the Ligand Root Mean Square Deviation (L-RMSD) to assess the predictive quality of ligand conformations. Meanwhile, the evaluation of side chain conformations' predictive quality is based on the side chain Root Mean Square Deviation (sc-RMSD). Let $X_l^{pred}$ represents the generated ligand pose, and $X_l^{gt}$ denotes the native ligand pose.

**3.2.1 L-RMSD.** We take into account Ligand Root Mean Square Deviation (L-RMSD) corrected for symmetry. The precise calculation formula is given below. Herein, $N$ represents the number of heavy atoms in the ligand, and isom denotes the isomers of the ligand molecular graph.

$$\text{L-RMSD} = \text{argmin}_{X_l^{\text{isom}} \sim \text{isom}\left(X_l^{\text{gt}}\right)} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_l^{\text{isom}}(i) - X_l^{\text{pred}}(i)\right)^2} \tag{11}$$

**3.2.2 Success rate.** L-RMSD < 2 Å is widely recognized as a benchmark indication of successful docking.[74] In fact, for cross-docking evaluations, the threshold for determining docking success can be relaxed to 2.5 Å. Nonetheless, to ensure equitable comparison, we adhere to the stricter threshold in this context.

**3.2.3 PB-success rate.** The PoseBusters test suite serves as a rigorous validation tool, assessing both the chemical and geometric consistency of a ligand, inclusive of its stereochemistry. Moreover, it evaluates the physical plausibility of intramolecular and intermolecular measurements, focusing on factors like the planarity of aromatic rings, canonical bond lengths, and potential protein–ligand clashes. Therefore, the PoseBusters suite provides users a more accurate and realistic estimation of the success rate, PB-success rate, through further checking the physical plausibility of poses with L-RMSD < 2 Å.

**3.2.4 sc-RMSD.** Given that the pocket backbone remains fixed, we compute the RMSD for the side chains of each residue individually and subsequently average the results. Furthermore, in consideration of the symmetrical topology inherent in side chain structures, symmetry corrections have been implemented for the ASP, GLU, PHE, and TYP residues. We regard an sc-RMSD value of less than 1 Å as indicative of success.

# 4 Results and discussion

## 4.1 Performance on the PDBbind time-split test set

The performance of DiffBindFR is first assessed on the PDBbind time-split test set.[46,67] We employed two metrics, including ligand Root Mean Square Deviation (L-RMSD) and side chain Root Mean Square Deviation (sc-RMSD) for the flexible docking evaluation. As is depicted from Fig. 2(B) (DiffBindFR-best),

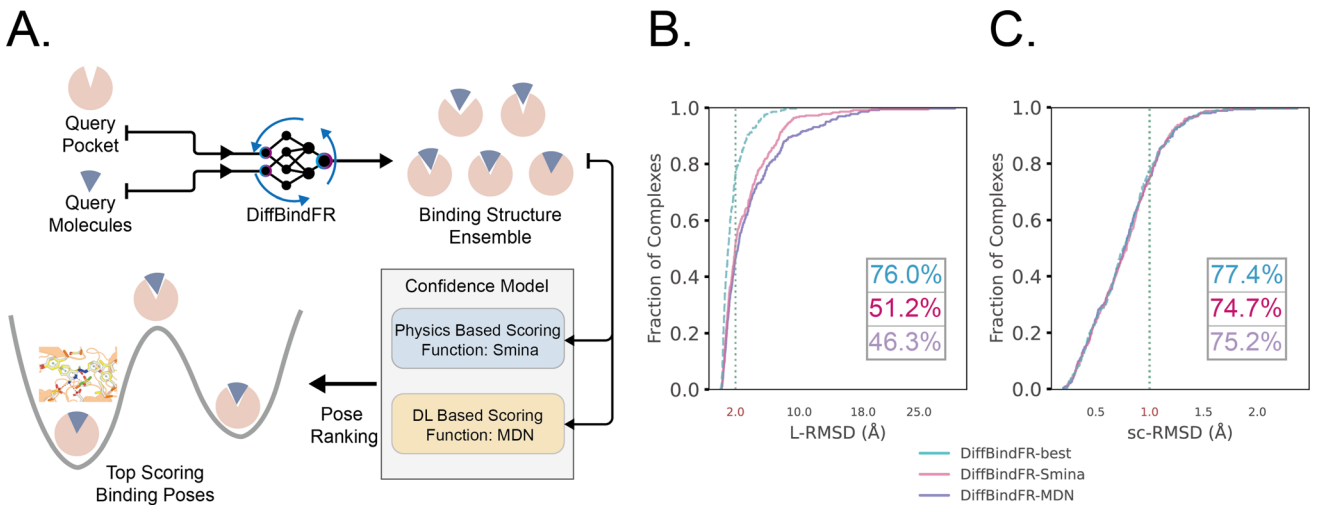



Fig. 2 (A) Overview of the DiffBindFR workflow. Various complex poses are generated by DiffBindFR network and confidence models are utilized to select the top-1 complex pose. Performance of DiffBindFR in PDBbind time-split test set for L-RMSD (B) and sc-RMSD (C). For each complex, 40 poses are generated. Distributions of L-RMSD and sc-RMSD are computed between DiffBindFR generated poses and ground-truth complex poses. Here, "DiffBindFR-best" means certain metrics are from the pose with the lowest L-RMSD generated by DiffBindFR model; "DiffBindFR-Smina" represents the DiffBindFR generated top-1 poses for each complex ranked by Smina scoring function; "DiffBindFR-MDN" represents the DiffBindFR generated top-1 poses for each complex ranked by MDN confidence model.

among the best poses (with lowest L-RMSD) from 40 DiffBindFR-generated poses for each complex, 76.0% of the ligand poses achieve successful docking (L-RMSD < 2 Å). Of these successful DiffBindFR-best poses, 77.4% exhibit reliable side chain recovery (sc-RMSD < 1 Å). The results of ablation experiments conducted on the hyperparameters related to network sampling and denoising within the DiffBindFR framework are detailed in ESI Fig. S3.† DiffBindFR-best represents the optimal scenario achievable by the DiffBindFR model, assuming a perfect confidence model could identify all the best poses. To enhance pose selection, we have developed two confidence models for ranking the generated poses (Fig. 2(A)). The first model (DiffBindFR-Smina) employs the all-atom physics-based scoring function Smina,[16] and the second one (DiffBindFR-MDN) utilizes a MDN (Mixture Density Network) network (Fig. S2†) trained on the PDBbind time-split training set by this work. Using Smina to select the top-1 binding poses, DiffBindFR-Smina attains a success rate of 51.2% in the PDBbind time-split test set. Among the top-1 poses ranked by Smina, 74.7% have reliable side chain recovery. The results demonstrate that DiffBindFR can accurately reconstruct the side chain conformations consistent with experimental pocket–ligand interactions, allowing the Smina to effectively select high-quality binding poses. When ranking sampled poses *via* the MDN model, DiffBindFR-MDN achieves a little bit lower success rate of 46.3% compared to DiffBindFR-Smina. Among the top-1 poses ranked by MDN confidence model, 75.2% have reliable side chain recovery.

Subsequently, the performance of traditional and recent deep learning-based methods is evaluated for comparison. The top-1 docking poses of each method, selected based on its confidence model or scoring function, are analyzed. DiffBindFR-Smina and DiffBindFR-MDN significantly outperform other deep learning-based pocket docking methods (Fig. 3(A)), including KarmaDock with RDKit[55] ligand conformation alignment (KarmaDock Align) and TankBind with predefined pocket (TankBind-Pocket). This highlights the advantage of our full-atom based model. Even when compared to the traditional rigid receptor docking methods AutoDock Vina and Glide, with the experimentally determined side chain conformations (redock), DiffBindFR-Smina achieves a marginally higher success rate without knowing the side chain conformations in the complex. It is noteworthy that on this test set, the re-docking success rates of Glide and AutoDock Vina are only 42.2% and 46.3%, respectively, which may be lower than the expectations for conventional methods in re-docking performance.[15,75] Our analysis suggests that this discrepancy could stem from dataset differences and the extent of exhaustive sampling (ESI Section 10†).

We further used Rosetta[76] to repack side chain conformations in these Holo structures to simulate an Apo-like state for each target protein. The docking success rate of Vina and Glide significantly decreases in these Apo-like proteins, underscoring the limitations of rigid receptor docking methods in handling side chain movements and the importance of flexible docking in virtual screening. To illustrate the challenges of flexible docking, we re-trained the KarmaDock by integrating a ResNet module[77] (ESI Fig. S4†) for predicting side chain torsion angles, resulting in a new model named KarmaDock-sc Align. DiffBindFR significantly surpasses KarmaDock-sc Align in terms of side chain recovery (Fig. 3(B)). Compared to KarmaDock, the performance of KarmaDock-sc Align significantly declines (Fig. 3(A)) due to the difficulty in balancing ligand coordinate recovery with side chain torsion recovery, highlighting the complexity inherent in flexible docking. As is widely recognized, factors like the number of heavy atoms and rotatable bonds in a ligand profoundly impact the success rate of conventional

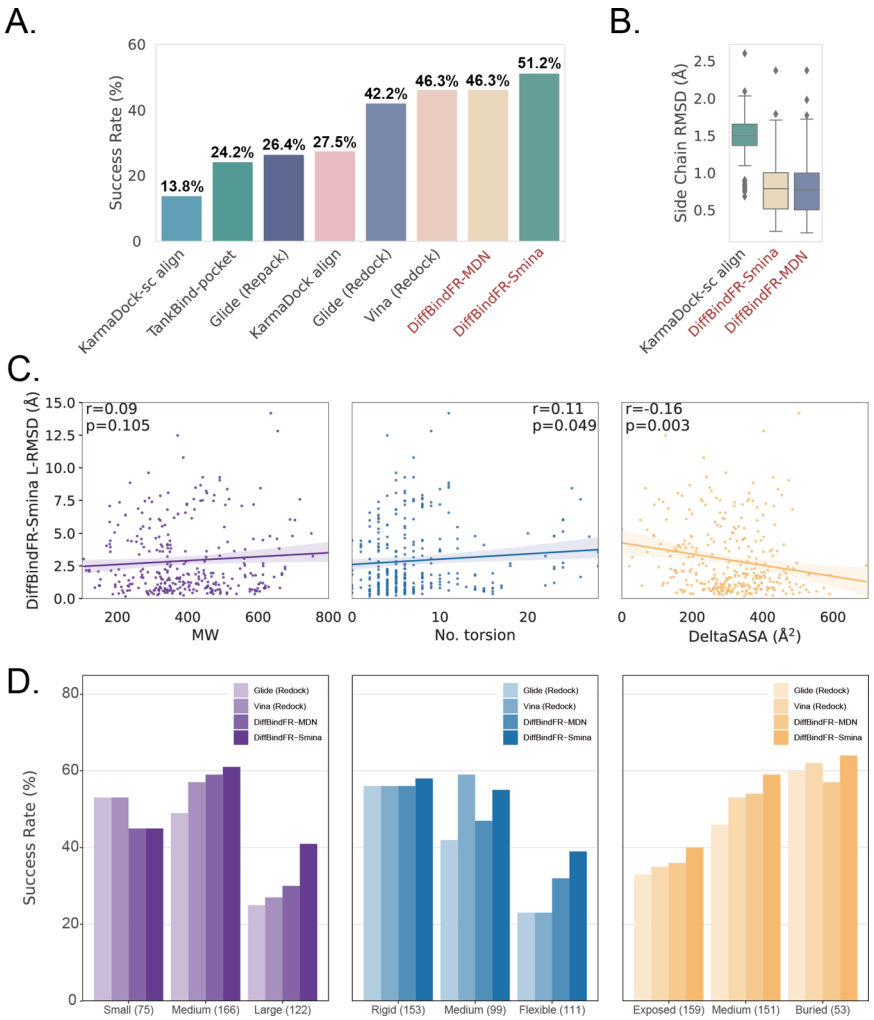



Fig. 3 (A) Success rate of various pocket docking methods. (B) The distribution of sc-RMSD of KarmaDock-sc Align, DiffBindFR-MDN and DiffBindFR-Smina. (C) The correlation between L-RMSD from DiffBindFR-Smina and molecular weight (MW), number of rotatable ligand bonds (no. torsion) and variation of solvent accessible surface area caused by binding (DeltaSASA). All three panels share the same *Y*-axis. (D) The impact of ligand molecular weight (small: ≤300 Da; medium: 300–500 Da; large: >500 Da), number of ligand rotatable bonds, (rigid: ≤5 bonds; medium: 5–9; flexible: ≥10) and DeltaSASA (exposed: ≤286 Å$^2$; medium: 286–418 Å$^2$; buried: >418 Å$^2$) on the docking success rates for Glide (Redock), Vina (Redock), DiffBind-Smina and DiffBind-MDN. All three panels share the same *Y*-axis.

docking programs.[78] Hence, we examine the relationship between L-RMSD from the DiffBindFR model and various ligand characteristics, such as molecular weight, rotatable torsion bonds, and changes in solvent accessible surface area (DeltaSASA) upon binding. Contrary to traditional methods,[14,17,31,47] as the difficulty of docking increases, including the increase in molecular weight, the number of rotatable bonds, and the decrease in the ligand buriedness, the L-RMSD achieved by DiffBindFR-Smina is not significantly affected (Fig. 3(C)). Additionally, DiffBindFR is able to achieve a more significant advantage in docking success rates compared to traditional methods (Fig. 3(D)).

### 4.2 Performance on the Posebusters test set

Given that the similarity between samples to be predicted and those used in training can influence the performance of deep learning methods, Buttenschoen *et al.*,[54] aiming for a more equitable comparison with traditional docking methods based on scoring functions, have curated a dataset called Posebusters test set from the PDB database. The Posebusters test set exclusively comprises 428 complexes on which the deep learning methods have not been trained. To evaluate the physical plausibility of poses generated by DiffBindFR, we compared its performance to other baseline methods using the Posebusters test set and the Posebusters suite,[54] a tool designed to assess the validity of ligand–protein complexes based on criteria including bond length, planarity of aromatic rings in ligands, and clashes between ligands and proteins. Success in docking is redefined as a pose having an L-RMSD less than 2 Å and simultaneously passing the physical validity check by Posebusters, with this success rate termed as the PB-success rate. The Posebusters test set comprises 428 distinct complexes released since 2021, with no overlap with the PDBbind v2020 dataset. To demonstrate that the success rate of DiffBindFR is not solely due to local

ligand energy relaxation and its superior in side chain packing, the performance of other methods is evaluated with stricter energy minimization for the ligand, given the experimental side chain conformation. For poses generated by DiffBindFR, ligand energy minimization is conducted using the side chains as predicted by the model. The energy minimization is performed using the AMBER ff14sb force field[79] for protein and the Sage force field[80] for ligand in OpenMM,[81] as used in the Posebusters paper.[54] Fig. 4(A) shows that traditional rigid receptor docking methods like Glide perform best on re-docking when provided with the correct Holo pocket environment, followed by Vina and Gold, with most of their generated docking poses being physically valid. However, their performance significantly deteriorates when docking with Rosetta-repacked proteins, highlighting their heavy dependence on side chain conformations. Traditional flexible docking methods rDock and VinaFlex are also involved in comparison. VinaFlex, heavily reliant on predefined flexible side chains, performs the worst in our scenario where information about flexible side chains is assumed unavailable. rDock, capable of optimizing functional groups prone to forming hydrogen bonds in side chains, achieves higher success rate in repacked proteins compared to Vina and Glide, but lower success rate in proteins with ground-truth side chains. For these traditional methods, their PB-success rate is only slightly lower than their overall success rate, indicating that most generated poses is validated by Posebusters suite due to the physical components in their scoring functions. Therefore, the post ligand optimization using force field does not cause obvious impact to their PB-success rate.

Among blind docking methods, DiffDock shows better performance (success rate of 38%) than TankBind (16%) and EquiBind (2%), but most of their generated poses are invalid due to ignoring protein side chains (Fig. 4(A)). Ligand energy minimization significantly improves the PB-success rate of DiffDock (35%). TankBind and EquiBind also see improvements in PB-success rate with energy minimization, but still lag behind DiffDock (Fig. 4(B)). Although blind docking is a tough task for its broad searching space in the whole protein, flexible pocket docking method like DiffBindFR, denoising a chaotic side chain conformation into a well-packed conformation having valid interaction with the ligand, has much more objectives for prediction. DiffBindFR, utilizing Smina scoring function or MDN network as the confidence model to select the top-1 pose from 40 generated ones, outperforms all other deep learning-based blind docking methods and pocket docking methods. DiffBindFR-Smina and DiffBindFR-MDN demonstrate both high success rate (50.2% for DiffBindFR-Smina and 48.1% for DiffBindFR-MDN) and PB-success rate (49.1% for DiffBindFR-Smina and 44.4% for DiffBindFR-MDN), with lower penalties by Posebusters compared to other deep learning-based methods, showcasing the capability of DiffBindFR in generating accurate and physically plausible complex poses. The performance of DiffBindFR is comparable to traditional rigid receptor docking methods using known ground-truth side chain conformations for redocking. As is depicted from ESI Fig. S7,† DiffBindFR shows its effectiveness in binding site identification and pocket side chain recovery on Posebusters test set, as well. Force field optimization has minimal impact on DiffBindFR generated structures, which also demonstrates the high physical plausibility of DiffBindFR generated poses. Among other deep learning-based pocket docking methods, KarmaDock Align achieves the highest success rate (30.4%) but a very low PB-success rate (6.1%). Force field optimization of ligands rescues most poses with L-RMSD < 2 Å into good

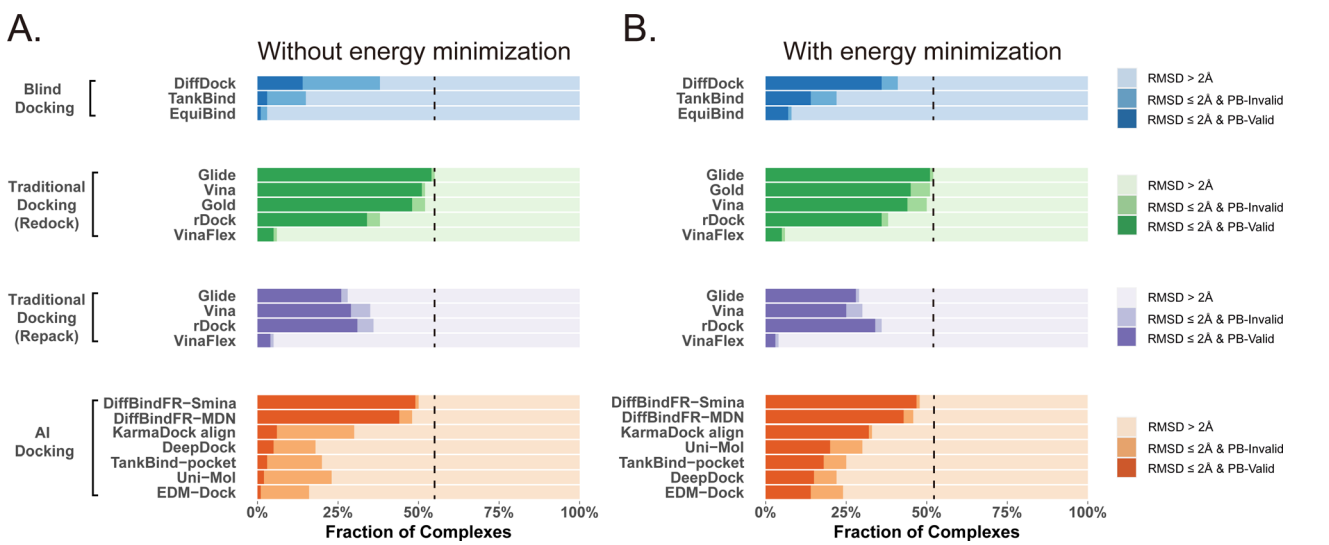



Fig. 4 Comparison of DiffBindFR with other methods boosted by force field optimization in Posebusters test set. The left column (A) shows the performance of methods without energy minimization, and the right one (B) shows the performance of methods with energy minimization. The lightest color represents the failure rate for docking, the moderate color represents the success rate, and the darkest color represent the PB-success rate. The blue color, green color, purple color and orange color represents performance of blind docking methods, traditional docking methods in redocking, traditional docking methods using repacked target proteins as input and deep learning-based docking methods, respectively. The dashed lines indicate the best-performing L-RMSD success rate (<2 Å) among all compared methods.

physical validity. KarmaDock Align, EDM-Dock, TankBind-pocket, Uni-Mol, and EDM-Dock, which focus on fitting the RMSD of the ligand during training, tend to ignore the intra energy of the generated poses and protein side chains, as is shown from ESI Fig. S9.† Indeed, force field optimization is not allowed in realistic docking to meet the demands of high-throughput screening.

Four specific cases from the Posebusters test set (ESI Fig. S10†), never trained or seen by DiffBindFR, are presented to highlight its superiority over other methods focusing solely on ligand coordinates while neglecting ligand conformation validity. In complexes with PDB ID 6TW5, 7PRM, 7T1D, and 7CD9, DiffBindFR successfully docks ligands into precise positions with valid conformations and recovers pocket side chains into good interaction with ligands. In contrast, KarmaDock Align, EDM-Dock, and TankBind-pocket fail to predict correct binding ligand poses, and their generated poses cannot pass the physical plausibility check of the Posebusters suite. As is shown in ESI Table S5,† ligand poses generated by EDM-Dock and TankBind-pocket exhibit both internal invalidity (including internal steric clash, bump aromatic ring, *etc.*) and steric clash with proteins, while KarmaDock Align, due to using RDKit for ligand pose alignment, frequently fails in reducing ligand–protein clash.

### 4.3 Performance on the CD cross-dock test set

To showcase the exceptional capabilities of DiffBindFR in flexible docking, we evaluate its performance in the more challenging task of cross-docking. We use a self-curated benchmark called the CD test set, which includes various cross-docking scenarios such as Apo–Holo and cross-docking between different Holo states with various protein families. CD test set contains 7 subsets, ApoRef [24], CASF2016 [67] with target proteins in the Apo and Holo states, GPCR-AF2 set with Apo-like proteins predicted by AlphaFold2,[29] DUDE27-HoloEns set with different protein Holo states and Ensemble sets featuring prominent docking targets including CDK2, EGFR and FXA. Cα RMSD of binding site backbone (within 5 Å cutoff away from crystal ligand) conformational changes in these subsets predominantly range between 0–2 Å, as shown in ESI Fig. S1.†

In these subsets, DiffBindFR-MDN and DiffBindFR-Smina achieve significantly higher PB-success rate (Table 1) than all the traditional docking methods and deep learning-based docking methods. When considering L-RMSD alone (Fig. 5), traditional rigid receptor docking methods such as Vina, Smina, LinF9,[82] and Glide underperform, were compared to deep learning-based methods that use main-chain coarse-grained representations of proteins. As indicated in Fig. 5 and ESI Table S6,† the L-RMSD median for methods like TankBind-pocket, EDM-Dock, and KarmaDock Align hovers around 2 Å across the subsets, whereas for traditional rigid receptor docking methods, it even surpasses 5 Å in subsets like CDK2, EGFR, and ApoRef. However, when physical plausibility is taken into account, the PB-success rate for TankBind-pocket, EDM-Dock, and KarmaDock Align drops to levels similar to traditional rigid receptor docking methods (below 10%). Notably, in the

Table 1 Success rate of various methods on CD cross-dock test set[a]

| Method | | PB-success rate | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Ensemble-CDK2 | Ensemble-EGFR | Ensemble-FXA | ApoRef | CASF2016 | DUDE27-HoloEns | GPCR-AF2 |
| Traditional rigid receptor docking methods | Vina | 0.079 | 0.060 | 0.344 | 0.082 | 0.294 | 0.160 | 0.136 |
| | LinF9 | 0.061 | 0.119 | 0.365 | 0.089 | 0.272 | 0.175 | 0.076 |
| | Smina | 0.079 | 0.090 | 0.362 | 0.077 | 0.304 | 0.157 | 0.152 |
| | Gnina | 0.099 | 0.090 | 0.394 | 0.082 | 0.320 | 0.179 | 0.106 |
| | Glide | 0.154 | 0.090 | 0.271 | 0.091 | 0.219 | 0.172 | 0.182 |
| Traditional flexible docking methods | VinaFlex | 0.013 | 0.000 | 0.005 | 0.015 | 0.023 | 0.034 | 0.045 |
| | rDock | 0.257 | 0.134 | 0.440 | 0.157 | 0.296 | 0.220 | 0.212 |
| Deep learning-based docking methods | TankBind-pocket | 0.100 | 0.015 | 0.110 | 0.040 | 0.123 | 0.049 | 0.015 |
| | EDM-Dock | 0.051 | 0.015 | 0.009 | 0.011 | 0.064 | 0.022 | 0.00 |
| | KarmaDock Align | 0.135 | 0.045 | 0.009 | 0.047 | 0.136 | 0.034 | 0.015 |
| | DiffBindFR-Smina | 0.564 | 0.403 | 0.789 | 0.434 | 0.566 | 0.295 | **0.227** |
| | DiffBindFR-MDN | **0.674** | **0.478** | **0.794** | **0.476** | **0.636** | **0.362** | 0.212 |

[a] Best performance in bold for the highest PB-success rate.

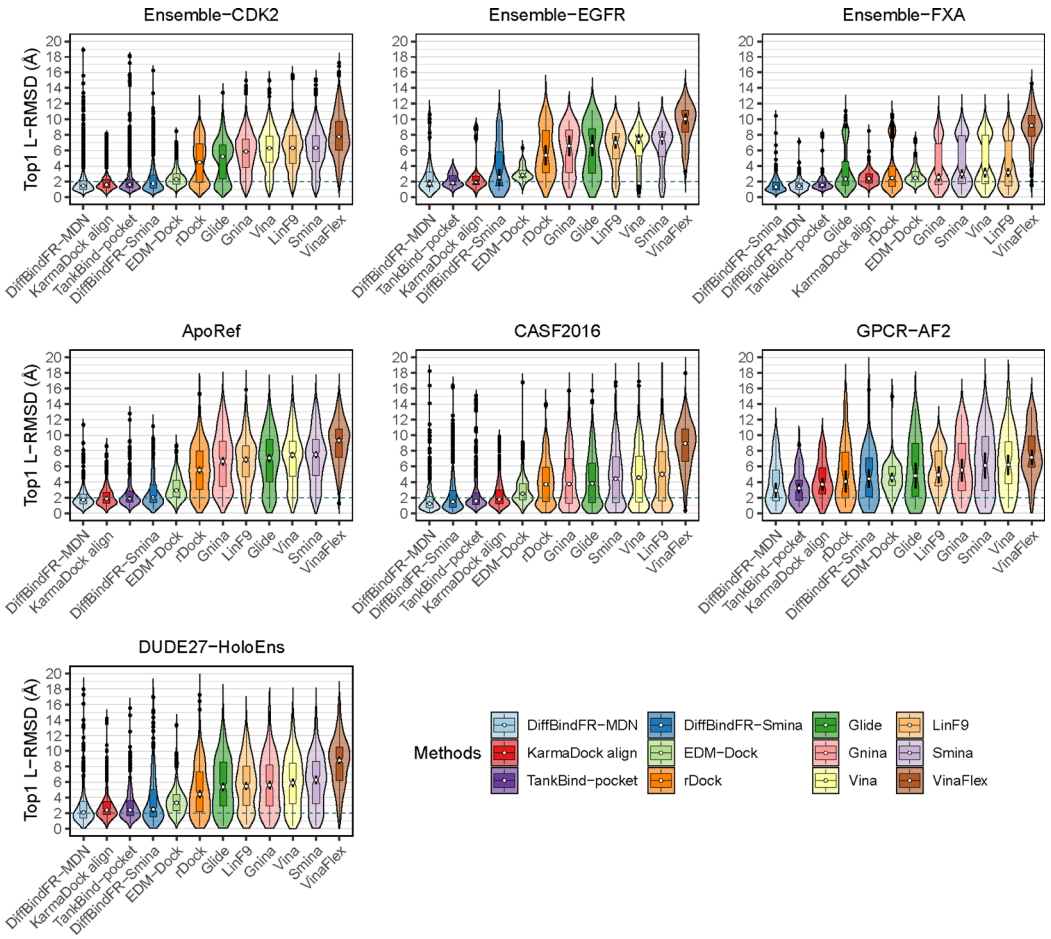



Fig. 5 The L-RMSD distribution of various methods on CD cross-dock test set. The median values are indicated by the white point inside the violin boxes, while the notches represent the 95% confidence interval around the median. Any outliers are shown as black points. The dashed lines indicate the L-RMSD criterion for determining success.

Ensemble-FXA subset, methods such as Vina, LinF9, Smina, Gnina, and Glide perform significantly better, achieving PB-success rates of 34.4%, 36.5%, 36.2%, 39.4%, and 27.1%, respectively. Traditional flexible methods such as VinaFlex and rDock, developed for cross-dock scenarios, were also evaluated. VinaFlex shows poorer performance than rigid receptor docking methods in both L-RMSD distribution and PB-success rate, due to its reliance on predefined side chains and limitations on the number of flexible side chains. The flexible method rDock, capable of optimizing side chains conducive to hydrogen bonding, outperforms all traditional rigid receptor docking methods in L-RMSD distribution and has higher PB-success rate than TankBind-pocket, EDM-Dock, and KarmaDock Align. We observed that both traditional rigid receptor and flexible docking methods perform better in subsets like Ensemble-FXA, DUDE27-HoloEns and CASF2016, where pocket backbone conformational changes are minimal (mostly between 0–0.5 Å, as is depicted from ESI Fig. S1†). Our method DiffBindFR, leveraging a full-atom based neural network to learn additional side chain movements, marginally outperforms all other methods in accurately recalling ligand coordinates and ensuring the validity of complex poses. On the CD test set, the MDN network surpasses the Smina scoring function for pose ranking. DiffBindFR-MDN achieves state-of-the-art PB-success rate of 67.4%, 47.8%, 79.4%, 47.6%, 36.2% and 63.6% in CDK2, EGFR, FXA, ApoRef, DUDE27-HoloEns and CASF2016, respectively. Furthermore, to investigate the efficacy of various methods in the context of cross-docking scenarios involving Apo–Holo pairs, we conduct a detailed computational assessment of these methods using a subset of 660 Apo–Holo pairs from the CASF2016 dataset. In this analysis, DiffBindFR-MDN emerges as the most effective technique, demonstrating superior performance in both terms of L-RMSD distribution and PB-success rate (56.1%, as is shown from ESI Table S7†). When the Holo structures for targets of interest are available, experts prefer to use these structures for virtual screening and lead optimization through molecular docking. Thus, we have constructed the DUDE27-HoloEns test set, entirely made up of Holo–Holo pairs, to evaluate the application potential of DiffBindFR in scenarios where target proteins possess resolved Holo structures. The results show that DiffBindFR-MDN performs the best on the DUDE27-HoloEns subset, achieving a PB-success rate of 36.2% and a median L-RMSD of 2.08 Å (ESI Table S6†), with DiffBind-Smina following closely (PB-success

rate of 29.5%). We find that traditional docking programs demonstrate enhanced proficiency in docking with Holo-state structures and perform better on the DUDE27-HoloEns subset and CASF2016 subset than other Apo–Holo cross-dock sets (ESI Tables S6–S9†), which could be attributed to subtle variations in conformational heterogeneity, favoring the binding of various bioactive molecules (see ESI Fig. S1 and Table S2†). Additionally, in extreme scenarios where the target protein lacks resolved Holo structures or even available crystal Apo structures, drug developers must rely on AlphaFold2 to obtain a predicted target structure as the starting point for drug design. On the GPCR-AF2 subset, serving as approximations of Apo–Holo cross-docking where the target proteins are predicted by AlphaFold2, DiffBindFR-MDN shows a slightly lower PB-success rate of 21.7% compared to DiffBindFR-Smina (22.7%). This decrease in performance is attributed to higher penalties from ligand–protein clashes.

Here, we present the docking poses of various methods on two examples from the GPCR-AF2 subset (Fig. 6). The first example involves a crystal structure with the PDB ID 6RZ6, where we investigate the target protein from 6RZ6 identified as the human Cysteinyl leukotriene receptor 2.[83] This receptor plays a role in regulating pro-inflammatory responses associated with allergic disorders. The ligand in this case is an antagonist, ONO-257036. We predict the AF2 structure (Apo-like) of the receptor protein and then dock the ligand molecule into this AF2 structure. Following binding sites alignment, the predicted AF2 structure exhibited a binding site (within 5 Å cutoff away from crystal ligand) RMSD of 0.73 Å when compared with the crystal structure. However, in the Apo-like AF2 predicted structure, residues R267 and L190 is found to block ONO-

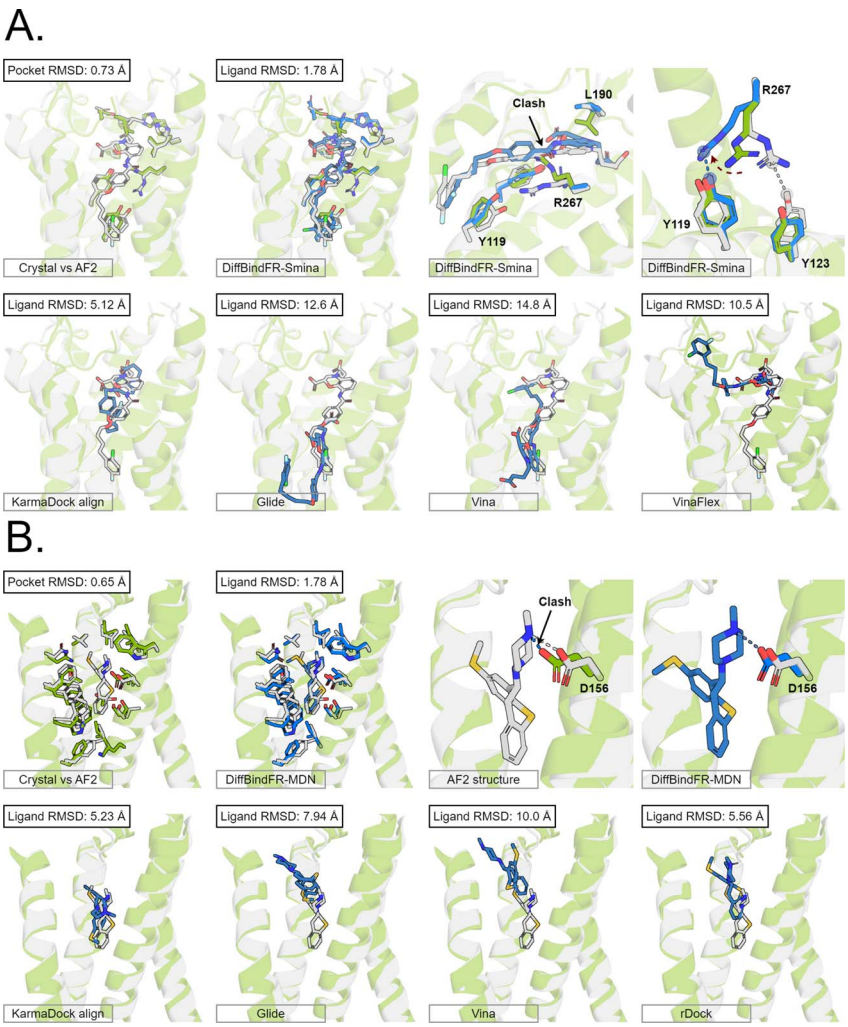



Fig. 6 The binding poses of two cases from the GPCR-AF2 subset in the CD test set. In all panels, Holo protein and ligand are shown in grey. AF2 modeled protein structure is shown in pale green. DiffBindFR sampled ligand and pocket side chains are shown in blue. Note that DiffBindFR calculations were done using the AF2 modeled backbone structures. For each frame, the docking method used is given at the bottom left. The pocket Cα RMSD between Holo and AF2 structure within 5 Å cutoff away from crystal ligand, and ligand RMSD are reported on left top, except the two right top frames in each panel for the comparison between AF2 modeled and DiffBindFR sampled side chain conformations. (A) Human cysteinyl leukotriene receptor 2 bound to its antagonist ONO-2570366 (PDB ID: 6RZ6). The side chains of Y119, Y123, L190 and R267 are shown; (B) 5-hydroxytryptamine receptor 2A bound to its inverse agonist methiothepin (PDB ID: 6WH4). The side chain of D155 is shown.

25703 binding to pocket, preventing traditional rigid receptor docking methods like Glide and Vina from locating the correct binding site. VinaFlex, despite its ability to leverage side chain flexibility, also fails to dock the ligand correctly. KarmaDock Align, similarly did not achieve correct docking, although it shows better L-RMSD than the methods mentioned above. In contrast, our flexible docking method, DiffBindFR, coupled with Smina and MDN confidence model can select the top-1 complex pose from the 40 poses. DiffBindFR adeptly repacks the side chain of the R267 residue, enabling the ligand to successfully dock into the correct binding position. The position of the R267 side chain predicted by DiffBindFR is somewhat different. In the crystal structure, the R267 side chain forms a hydrogen bond interaction with the Y123 side chain, whereas in the DiffBindFR-predicted structure, R267 forms a hydrogen bond interaction with the Y119 side chain, which can be attributed to the similarity in the spatial positions of the Y123 residue and the Y119 residue relative to R267.

The second example involves a crystal structure with the PDB ID 6WH4, where the target protein is the human 5-HT2A serotonin receptor, which is associated with the actions mediated by psychedelics,[84] and the ligand is methiothepin. Following alignment of the binding sites, the AF2-predicted structure displays a pocket RMSD of 0.65 Å when compared to the crystal structure. In the AF2 structure, the side chain of the D155 residue has vdW (van der Waals) clash (pair distance is 1.5 Å) with the ligand, leading to the failure of docking attempts by Glide and Vina. Although KarmaDock Align and the traditional flexible docking method rDock predicts the ligand position with a smaller L-RMSD, they still do not achieve successful docking. In contrast, the top-1 complex pose selected by the MDN confidence model in DiffBindFR not only accurately reproduces the ligand position but also successfully repacks the side chain of D155 residue, enabling it to form electrostatic interactions with the ligand. This example further demonstrates the capability of DiffBindFR to effectively manage protein–ligand interactions, particularly in challenging docking scenarios.

We also present four cases from ApoRef subset where DiffBindFR successfully docks ligands into the Apo pockets of crystal structures (ESI Fig. S11†), with these complexes not having appeared in the training set. The PDB IDs for these four cross-dock examples are as follows: Holo: 1ZGY, Apo: 1PRG; Holo: 2XIR, Apo: 1VR2; Holo: 3UVR, Apo: 1WFC; and Holo: 3RM6, Apo: 4EK3. In each of these instances, top-1 complex poses generated by DiffBindFR accurately recover the ligand poses, while side chains in the Apo state pockets that would otherwise clash with the ligand are also optimized. These results highlight the potential of DiffBindFR in aiding researchers to study detailed interactions in real scenarios when no complex structures are available and provide insights for further lead optimization.

The superior performance of DiffBindFR on different Apo–Holo, Holo–Holo and AF2 predicted-Holo cross-docking subsets have shown that the potential of DiffBindFR under the real circumstances for application.

### 4.4 Performance on the DUDE27-AF2 induced fit docking test set

Aside from rDock and VinaFlex, to explicitly consider the flexibility of the protein pocket, Miller *et al.*[38] developed an induced fit docking workflow named IFD-MD, which integrates molecular docking with molecular dynamics to simulate the induced-fit effects during binding. In this docking experiment, we utilize the dataset by Zhang *et al.*[26] as the test set (hereafter referred to as the DUDE27-AF2 test set), which comprises the Holo and AF2-predicted structures for 27 targets. This dataset also includes docking results based on the IFD-MD refined AF2 predicted structures. On the DUDE27-AF2 test set, we have compared the performance of DiffBindFR with other flexible docking methods, including VinaFlex, rDock, and IFD-MD, in

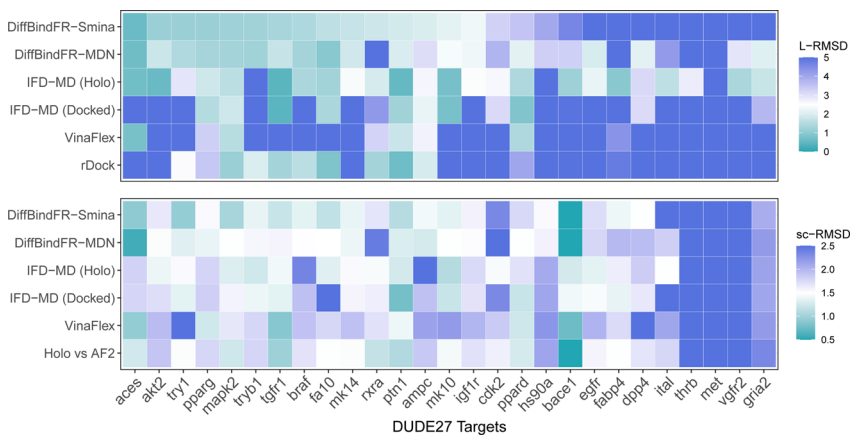



Fig. 7 The L-RMSD and sc-RMSD performance of various methods on the DUDE27-AF2 subset. Here, sc-RMSD refers to the RMSD of flexible side chains as specified based on expert experience. IFD-MD (Holo) and IFD-MD (Docked) represent the approaches where IFD-MD uses the aligned Holo crystal ligand pose as the template and the ligand pose generated by docking into the pocket of AF2 predicted structure using Glide, respectively. "Holo *vs.* AF2" refers to the sc-RMSD between the Holo pocket and the AF2 predicted protein pocket. Values exceeding the upper limit will be assigned the color designated for the upper limit (5 Å for L-RMSD and 2.5 Å for sc-RMSD).

their ability to accurately dock ligands into the correct poses while optimizing the AF2 predicted protein structures. IFD-MD depends on a ligand pose template to further optimize protein conformations. The DUDE27-AF2 test set provides two types of ligand pose templates. The first one uses the aligned crystal ligand pose as the template to induce the protein conformation, which can be seen as the upper limit of IFD-MD's capability since the crystal ligand binding pose cannot be known in advance. The more realistic approach docks the Holo ligand into the AF2 structures with Glide and uses the resulting pose as the template for IFD-MD. By examining the aligned structures of the AF2 predicted and Holo structures, we select flexible residues based on expert experience to provide to VinaFlex as input and also to evaluate each method's ability of optimizing protein structures. As the results shown in Fig. 7 indicate, generally, the greater the sc-RMSD difference between the AF2 predicted structures and the Holo structures, the more challenging it is for the methods to accurately dock the ligand pose and recover side chain conformation. As expected, because the ground-truth ligand pose template is provided in advance, IFD-MD using the crystal ligand binding pose as a template performed better than other methods, maintaining 59% of ligand poses within an L-RMSD of 2 Å, with a median RMSD of 1.74 Å (ESI Table S11†). However, the real ligand binding positions cannot be obtained beforehand. Compared to IFD-MD using Glide generated ligand poses by docking the ligand into AF2 predicted pocket as a template, VinaFlex, and rDock, DiffBindFR achieves better results in predicting the poses of the ligands. Specifically, the median L-RMSD for DiffBind-Smina and DiffBind-MDN are 2.12 Å and 2.15 Å, respectively (ESI Table S11†). In terms of side chain optimization, whether for flexible side chains selected based on expert experience or for the overall pocket side chains, DiffBindFR demonstrates superior side chain optimization performance, even surpassing that of IFD-MD using the ground-truth ligand pose as a template (ESI Table S12†). This showcases the exceptional capability of DiffBindFR in flexible docking and its potential application on AF2 predicted target protein structures.

# 5 Conclusions

In this research, we have developed a full-atom diffusion model, DiffBindFR, for flexible pocket docking. DiffBindFR is capable of explicitly simulating the interactions of full atoms between the pocket side chains and the ligand molecules, which is extremely hard for previous docking methodologies. Our method not only surpasses traditional approaches in terms of the docking success rate but also achieves state-of-the-art (SOTA) levels in generating plausible docking conformations when compared to recent deep learning methods, as evidenced by evaluations conducted on the PDBbind and Posebusters test sets. Furthermore, starting from a random side chain conformation, DiffBindFR can accurately dock molecules while concurrently recovering the side chain conformations.

On the cross-docking benchmark, CD test set, DiffBindFR has also demonstrated superior performance. Notably, previous methods that employ deep learning to characterize protein pockets through coarse-grained main-chain representations also show promise results, but they lead to a lack of detailed information regarding the interactions between side chain atoms and ligands. DiffBindFR that simulates side chains addresses this gap in deep learning methodologies. Previous research[57–59] has indicated that the majority of proteins undergo minimal backbone alterations upon ligand binding, with the primary conformational changes caused by side chains. Therefore, DiffBindFR remains adept at predicting accurate docking poses in most scenarios with slight protein backbone movements. However, in cases where the pocket backbone exhibits significant flexibility (such as cryptic sites[85]), the docking results may not be as satisfactory. Recent benchmarking studies[25–27,30,86,87] have revealed that structures modeled by AF2 tend to exhibit a pocket backbone conformation more akin to the Holo state, presumably owing to the conservative functional sites within the multiple sequence alignment (MSA). However, inaccuracies in side chain placement often lead to suboptimal virtual screening performance when compared with the Holo pocket. In this context, DiffBindFR emerges as a promising tool for refining the side chains in AF2 modeled structures, potentially enhancing enrichment in virtual screening campaigns in the absence of available Holo structures. Further investigation of this application will be in future research endeavors.

The physical validity of DiffBindFR generated complex poses, coupled with the simulated detailed three-dimensional interactions, provides users with correct interactions to facilitate further optimization. In addition, the conformation alterations predicted by DiffBindFR will significantly augment comprehension of the molecular mechanisms underlying specific actions, and better elucidate the relationship between structure and function.

# 6 Methods

## 6.1 Data representation

The complete set of heavy atoms from the ligand molecule and protein pocket is structured into a heterogeneous graph $G = (\nu, \mathcal{E})$, where each atom corresponds to a node. For the node representation $\nu_p$ of pocket residue atoms, we employ one-hot encoding encompassing atom type, residue type, and whether the atom is part of the backbone. The ligand node features $\nu_l$ include atom type, hybridization type, atomic connectivity, explicit valence, implicit valence, number of rings it belonging to, aromaticity, formal charge, partial charge, chirality type, the number of radical electrons, the number of hydrogens, and whether it is in an N-membered ring (with nitrogen ranging from 3 to 8). Furthermore, pharmacophore features such as hydrogen bond acceptor/donor, aromaticity, hydrophobicity, and positive/negative charge are integrated. The edges $e_{ij}$, based on the covalent bonds of ligand, incorporate features like bond type, stereochemistry, conjugation, and whether the bond is part of a ring system. Diffusion times $t$ are encoded using a sinusoidal format and are concatenated to the scalar representations of nodes and edges. For ligand atoms, internal edges $\mathcal{E}_{ll}$ connected by covalent bonds are pre-

constructed. For pocket atoms, in addition to covalent bonds, edges $\mathcal{E}_{pp}$ are linked between pocket atoms and their own $C_\alpha$ and $C_\beta$ atoms. During the forward inference of the model, edges are dynamically constructed based on the three-dimensional coordinates of all atoms. Within the ligand molecule, the graph construction uses a cutoff radius of 5 Å, and a similar cutoff is applied for the full atom graph of the pocket and directed edges from pocket to ligand atoms $\mathcal{E}_{pl}$. These edges, serving as non-covalent interactions, solely encode distances. Given the model's need to predict ligand translational updates, it's essential for the ligand to be aware of the entire pocket atom's position. Therefore, directed edges $\mathcal{E}_{lp}$ from ligand to pocket atoms are dynamically constructed based on the diffusion process, with the translational noise determining the cutoff radius as $0.2\sigma_T + 5$ Å. This ensures that even in high noise scenarios, where the ligand is distant from the pocket, there remains an informational interaction between the ligand and the pocket, thereby pulling the ligand closer. All edges from the heterogeneous graph are $\{\mathcal{E}_{ll}, \mathcal{E}_{pp}, \mathcal{E}_{pl}, \mathcal{E}_{lp}\}$, and their distance features utilize Gaussian radial basis for encoding.

### 6.2 DiffBindFR score network

The architecture of DiffBindFR is meticulously crafted upon the foundation provided by the e3nn library.[53] It primarily comprises the following pivotal components: a module for input embedding, modules for intra-molecular interaction encoding, and modules dedicated to inter-molecular interaction encoding. The network ingests a geometric heterogeneous graph, encompassing invariant scalar representations of both ligand heavy atom nodes $\nu_l$ and pocket residue atom nodes $\nu_p$. We harness the irreducible representations (Irreps) to encode features by spherical harmonics. As the depth of feature encoding advances, the scalar inputs evolve towards higher-order physical quantity representations. Every interaction module is constructed using the Tensor Product Layer (TPL), establishing SE(3) equivariant message-passing functions. The tensor products are realized by encoding edge vectors with spherical harmonic functions and then doing spherical tensor product of Irreps with path weights. The weights of these tensor products are derived from a transformation of node representations constituting the edge and the edge representation itself through a layer of Multi-Layer Perceptrons (MLP); these weights also constitute the primary learnable parameters at each layer. For any given submodule, the general formula for message passing to node a is:

$$h_a \leftarrow h_a \underset{z\in\{l,p\}}{\oplus} \mathrm{LN}^{(z_a,z)}\left(\frac{1}{\left|\mathbf{N}_a^{(z)}\right|}\sum_{b\in\mathbf{N}_a^{(z)}} Y(\hat{r}_{ab})\otimes\psi_{ab}\ h_b\right) \tag{12}$$

$h_a = (h_a^0, \overrightarrow{h_a})$ represents Irreps of node a, which is the concatenation of scalar representation $h_a^0$ and vector representation $\overrightarrow{h_a}$. $z_a$ is the node type of node a, and z can be any node type from the pocket node or the ligand node. $\mathbf{N}_a^{(z)}$ denotes the neighbour nodes of node a. $Y$ are the spherical harmonics up to $l = 2$. LN is the equivariant layer normalization. $\oplus$ refers to normal vector addition, and $\otimes\psi_{ab}$ refers to the spherical tensor product of Irreps with path weights, with $\psi_{ab} = \mathrm{MLP}^{(z_a,z)}(e_{ab}, h_a^0, h_b^0)$ following the graph message passing paradigm.

For predicting the scores of ligand translation and rotation, we construct a node o for each ligand center and gather the message from other ligand nodes to the center. We output the final single odd and single even vectors through layer normalization for translational and rotational score prediction as follows:

$$\left[h_l^{(1o)},\ h_l^{(1e)}\right] \leftarrow \frac{1}{|\nu_l|}\sum_{a\in\nu_l} Y(\hat{r}_{oa})\otimes\psi_{oa}h_a, \quad \text{with } \psi_{oa} = \mathrm{MLP}(\mu_{oa}, h_a^0) \tag{13}$$

$\mu_{oa}$ denotes the Gaussian radial embeddings for distance $r_{oa}$. $h_l^{(1o)}$ is the predicted score for translation. $h_l^{(1e)}$ is the predicted score for rotation axis $\hat{\omega}$.

For both the rotatable bonds of ligands and the dihedral angles of protein residue side chains, updates for each angle are anticipated based on a consistent paradigm. We define the central axis of the rotatable bond or dihedral angle as $B = (i, j)$, represented by a bond formed by atoms i and j. Further, the center of the rotatable bond is denoted as c. A radius graph of ligand nodes with a 4 Å cutoff is constructed to predict the torsion score of the ligand rotatable bonds.

$$h_c \leftarrow \frac{1}{\mathbf{N}_c}\sum_{a\in N_c} Y^2(\hat{r}_{ab})\otimes Y(\hat{r}_{ca})\otimes\psi_{ca}h_a, \quad \text{with } \psi_{ca} = \mathrm{MLP}(\mu_{ca}, h_a, h_i + h_j) \tag{14}$$

To satisfy the parity of dihedral angles, spherical harmonics $Y^2$ up to $l = 2$ is utilized. We will employ the scalar features derived from the final obtained $h_c$ to predict the torsion angles. An analogous procedure is adopted for the torsion of pocket side chains.

### 6.3 Model training

DiffBindFR neural network was trained using the AdamW optimizer[88] with a learning rate of 0.0005 and a batch size of 64 for 1000 epochs on eight 80 GB NVIDIA A800 TENSOR CORE GPUs. MDN confidence model was trained using the Adam optimizer[89] with a batch size of 256 for 1000 epochs on four 32 GB NVIDIA Tesla V100-SXM2 GPUs.

## Data availability

The protein–ligand complexes of PDBbind v2020 dataset were downloaded from **https://zenodo.org/records/6408497**. The protein–ligand complexes of Posebusters test set were downloaded from **https://zenodo.org/records/8278563**. The protein–ligand complexes of CD cross-dock test set are publicly available at **https://doi.org/10.5281/zenodo.10816044**.[90]

## Code availability

The source code of DiffBindFR is publicly available at **https://github.com/HBioquant/DiffBindFR**.

## Author contributions

J. Z. and Z. G. designed the research, wrote source code and performed the experiments. J. P. and L. L. designed and supervised the project. J. Z. analyzed the experimental results. Z. G. and J. Z wrote the manuscript. J. P. and L. L. revised the manuscript. All authors read and approved the final manuscript.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

This work has been supported in part by the National Natural Science Foundation of China (22033001 and 32270689), the National Key R&D Program of China (2023YFF1205103), the Chinese Academy of Medical Sciences (2021-I2M-5-014) and the Anhui's Plans for Major Provincial Science&Technology Projects (202303a07020009). We thank the Computing Platform of the Center for Life Science (Peking University) for providing resources for the GPU-based model training. Part of the computation was performed on the computing platform of the Infinite Intelligence Pharma Ltd.

## Notes and references

1 J. S. Handen, *Drug Discov. World*, 2002, 47–50.
2 L. M. Mayr and D. Bojanic, *Curr. Opin. Pharmacol.*, 2009, **9**, 580–588.
3 A. L. Satz, A. Brunschweiger, M. E. Flanagan, A. Gloger, N. J. Hansen, L. Kuai, V. B. Kunig, X. Lu, D. Madsen, L. A. Marcaurelle, *et al.*, *Nat. Rev. Methods Primers*, 2022, **2**, 3.
4 T. Fink, H. Bruggesser and J.-L. Reymond, *Angew. Chem., Int. Ed.*, 2005, **44**, 1504–1508.
5 C. Grebner, E. Malmerberg, A. Shewmaker, J. Batista, A. Nicholls and J. Sadowski, *J. Chem. Inf. Model.*, 2019, **60**, 4274–4282.
6 A. V. Sadybekov and V. Katritch, *Nature*, 2023, **616**, 673–685.
7 J. Lyu, J. J. Irwin and B. K. Shoichet, *Nat. Chem. Biol.*, 2023, **19**, 712–718.
8 C. Gorgulla, A. Boeszoermenyi, Z.-F. Wang, P. D. Fischer, P. W. Coote, K. M. Padmanabha Das, Y. S. Malets, D. S. Radchenko, Y. S. Moroz, D. A. Scott, *et al.*, *Nature*, 2020, **580**, 663–668.
9 H. Zhu, Y. Zhang, W. Li and N. Huang, *Int. J. Mol. Sci.*, 2022, **23**, 15961.
10 J. Lyu, S. Wang, T. E. Balius, I. Singh, A. Levit, Y. S. Moroz, M. J. O'Meara, T. Che, E. Algaa, K. Tolmachova, *et al.*, *Nature*, 2019, **566**, 224–229.
11 J. Fan, A. Fu and L. Zhang, *Quant. Biol.*, 2019, **7**, 83–89.
12 B. J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C. M. Webb, R. M. Stein, E. A. Fink, T. E. Balius, J. Carlsson, J. J. Irwin, *et al.*, *Nat. Protoc.*, 2021, **16**, 4799–4832.
13 G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell and A. J. Olson, *J. Comput. Chem.*, 2009, **30**, 2785–2791.
14 O. Trott and A. J. Olson, *J. Comput. Chem.*, 2010, **31**, 455–461.
15 J. Eberhardt, D. Santos-Martins, A. F. Tillack and S. Forli, *J. Chem. Inf. Model.*, 2021, **61**, 3891–3898.
16 D. R. Koes, M. P. Baumgartner and C. J. Camacho, *J. Chem. Inf. Model.*, 2013, **53**, 1893–1904.
17 R. A. Friesner, J. L. Banks, R. B. Murphy, T. A. Halgren, J. J. Klicic, D. T. Mainz, M. P. Repasky, E. H. Knoll, M. Shelley, J. K. Perry, *et al.*, *J. Med. Chem.*, 2004, **47**, 1739–1749.
18 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Taylor, *J. Mol. Biol.*, 1997, **267**, 727–748.
19 A. Lauria, M. Tutone and A. M. Almerico, *Eur. J. Med. Chem.*, 2011, **46**, 4274–4280.
20 Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian and T. Hou, *Phys. Chem. Chem. Phys.*, 2016, **18**, 12964–12975.
21 D. A. Antunes, D. Devaurs and L. E. Kavraki, *Expert Opin. Drug Discovery*, 2015, **10**, 1301–1313.
22 S. L. McGovern and B. K. Shoichet, *J. Med. Chem.*, 2003, **46**, 2895–2907.
23 H. S. Lee, C. S. Lee, J. S. Kim, D. H. Kim and H. Choe, *J. Chem. Inf. Model.*, 2009, **49**, 2419–2428.
24 J. Zhang, H. Li, X. Zhao, Q. Wu and S.-Y. Huang, *J. Chem. Inf. Model.*, 2022, **62**, 5806–5820.
25 A. M. Díaz-Rovira, H. Martin, T. Beuming, L. Díaz, V. Guallar and S. S. Ray, *J. Chem. Inf. Model.*, 2023, **63**, 1668–1674.
26 Y. Zhang, M. Vass, D. Shi, E. Abualrous, J. M. Chambers, N. Chopra, C. Higgs, K. Kasavajhala, H. Li, P. Nandekar, *et al.*, *J. Chem. Inf. Model.*, 2023, **63**, 1656–1667.
27 C. Kersten, S. Clower and F. Barthels, *J. Chem. Inf. Model.*, 2023, **63**, 2218–2225.
28 T. Scior, A. Bender, G. Tresadern, J. L. Medina-Franco, K. Martínez-Mayorga, T. Langer, K. Cuanalo-Contreras and D. K. Agrafiotis, *J. Chem. Inf. Model.*, 2012, **52**, 867–881.
29 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, *Nature*, 2021, **596**, 583–589.
30 M. Karelina, J. J. Noh and R. O. Dror, *eLife*, 2023, **12**, RP89386.
31 S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard and S. D. Morley, *PLoS Comput. Biol.*, 2014, **10**, e1003571.
32 P. A. Ravindranath, S. Forli, D. S. Goodsell, A. J. Olson and M. F. Sanner, *PLoS Comput. Biol.*, 2015, **11**, e1004586.
33 A. Basciu, L. Callea, S. Motta, A. M. Bonvin, L. Bonati and A. V. Vargiu, *Annu. Rep. Med. Chem.*, 2022, **59**, 43–97.
34 R. E. Amaro, J. Baudry, J. Chodera, Ö. Demir, J. A. McCammon, Y. Miao and J. C. Smith, *Biophys. J.*, 2018, **114**, 2271–2278.
35 M. De Vivo, M. Masetti, G. Bottegoni and A. Cavalli, *J. Med. Chem.*, 2016, **59**, 4035–4061.
36 A. Wang, Y. Zhang, H. Chu, C. Liao, Z. Zhang and G. Li, *J. Chem. Inf. Model.*, 2020, **60**, 2939–2950.
37 W. Evangelista Falcon, S. R. Ellingson, J. C. Smith and J. Baudry, *J. Phys. Chem. B*, 2019, **123**, 5189–5195.
38 E. B. Miller, R. B. Murphy, D. Sindhikara, K. W. Borrelli, M. J. Grisewood, F. Ranalli, S. L. Dixon, S. Jerome,

N. A. Boyles, T. Day, *et al.*, *J. Chem. Theory Comput.*, 2021, **17**, 2630–2639.
39 D. Coskun, M. Lihan, J. P. Rodrigues, M. Vass, D. Robinson, R. A. Friesner and E. B. Miller, *J. Chem. Theory Comput.*, 2023, **20**, 477–489.
40 Y. Yu, S. Lu, Z. Gao, H. Zheng and G. Ke, *ICLR 2023-Machine Learning for Drug Discovery Workshop*, 2023.
41 O. Méndez-Lucio, M. Ahmad, E. A. del Rio-Chanona and J. K. Wegner, *Nat. Mach. Intell.*, 2021, **3**, 1033–1039.
42 W. Lu, Q. Wu, J. Zhang, J. Rao, C. Li and S. Zheng, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 7236–7249.
43 M. R. Masters, A. H. Mahmoud, Y. Wei and M. A. Lill, *J. Chem. Inf. Model.*, 2023, **63**, 1695–1707.
44 V. G. Satorras, E. Hoogeboom and M. Welling, *International Conference on Machine Learning*, 2021, pp. 9323–9332.
45 K. Atz, F. Grisoni and G. Schneider, *Nat. Mach. Intell.*, 2021, **3**, 1023–1032.
46 H. Stärk, O. Ganea, L. Pattanaik, R. Barzilay and T. Jaakkola, *International Conference on Machine Learning*, 2022, pp. 20503–20521.
47 J. Zhang, K. He, T. Dong and J. Wu, *Research Square*, 2022, DOI: **10.21203/rs.3.rs-1454132/v1**.
48 Y. Zhang, H. Cai, C. Shi and J. Tang, *The Eleventh International Conference on Learning Representations*, 2022.
49 G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, L. Zhang and G. Ke, *The Eleventh International Conference on Learning Representations*, 2022.
50 X. Zhang, O. Zhang, C. Shen, W. Qu, S. Chen, H. Cao, Y. Kang, Z. Wang, E. Wang, J. Zhang, *et al.*, *Nat. Comput. Sci.*, 2023, **3**, 789–804.
51 G. Corso, H. Stärk, B. Jing, R. Barzilay and T. S. Jaakkola, *The Eleventh International Conference on Learning Representations*, 2022.
52 J. Ho, A. Jain and P. Abbeel, *Adv. Neural Inf. Process. Syst.*, 2020, **33**, 6840–6851.
53 M. Geiger and T. Smidt, *arXiv*, 2022, preprint, arXiv:2207.09453, DOI: **10.48550/arXiv.2207.09453**.
54 M. Buttenschoen, G. M. Morris and C. M. Deane, *Chem. Sci.*, 2024, **15**, 3130–3139.
55 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.
56 T. Dong, Z. Yang, J. Zhou and C. Y.-C. Chen, *J. Chem. Theory Comput.*, 2023, **19**, 8446–8459.
57 F. Gaudreault, M. Chartier and R. Najmanovich, *Bioinformatics*, 2012, **28**, i423–i430.
58 J. J. Clark, M. L. Benson, R. D. Smith and H. A. Carlson, *PLoS Comput. Biol.*, 2019, **15**, e1006705.
59 S. A. Wankowicz, S. H. de Oliveira, D. W. Hogan, H. van den Bedem and J. S. Fraser, *eLife*, 2022, **11**, e74114.
60 Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon and B. Poole, *The Tenth International Conference on Learning Representations*, 2021.
61 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, *International Conference on Machine Learning*, 2017, pp. 1263–1272.
62 B. D. Anderson, *Stoch. Process. Appl.*, 1982, **12**, 313–326.
63 D. I. Nikolayev, T. I. Savyolov, *et al.*, *Texture, Stress, Microstruct.*, 1997, **29**, 201–233.
64 A. Leach, S. M. Schmon, M. T. Degiacomi and C. G. Willcocks, *ICLR 2022 Workshop on Geometrical and Topological Representation Learning*, 2022.
65 B. Jing, G. Corso, J. Chang, R. Barzilay and T. Jaakkola, *Adv. Neural Inf. Process. Syst.*, 2022, **35**, 24240–24253.
66 E. Rodolà, Z. Lähner, A. M. Bronstein, M. M. Bronstein and J. Solomon, *Comput. Graph. Forum*, 2019, 678–689.
67 Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li and R. Wang, *Acc. Chem. Res.*, 2017, **50**, 302–309.
68 M. Volkov, J.-A. Turk, N. Drizard, N. Martin, B. Hoffmann, Y. Gaston-Mathé and D. Rognan, *J. Med. Chem.*, 2022, **65**, 7946–7958.
69 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
70 M. M. Mysinger, M. Carchia, J. J. Irwin and B. K. Shoichet, *J. Med. Chem.*, 2012, **55**, 6582–6594.
71 R. Aggarwal, A. Gupta and U. Priyakumar, *arXiv*, 2021, preprint, arXiv:2108.09926, DOI: **10.48550/arXiv.2108.09926**.
72 C. P. Feidakis, R. Krivak, D. Hoksza and M. Novotny, *Bioinformatics*, 2022, **38**, 5452–5453.
73 S. Bietz and M. Rarey, *J. Chem. Inf. Model.*, 2016, **56**, 248–259.
74 A. Alhossary, S. D. Handoko, Y. Mu and C.-K. Kwoh, *Bioinformatics*, 2015, **31**, 2214–2216.
75 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2018, **59**, 895–913.
76 C. A. Rohl, C. E. Strauss, K. M. Misura and D. Baker, *Methods in Enzymology*, Elsevier, 2004, vol. 383, pp. 66–93.
77 K. He, X. Zhang, S. Ren and J. Sun, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
78 N. S. Pagadala, K. Syed and J. Tuszynski, *Biophys. Rev.*, 2017, **9**, 91–102.
79 J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, *J. Chem. Theory Comput.*, 2015, **11**, 3696–3713.
80 S. Boothroyd, P. K. Behara, O. C. Madin, D. F. Hahn, H. Jang, V. Gapsys, J. R. Wagner, J. T. Horton, D. L. Dotson, M. W. Thompson, *et al.*, *J. Chem. Theory Comput.*, 2023, **19**, 3251–3275.
81 P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, *et al.*, *PLoS Comput. Biol.*, 2017, **13**, e1005659.
82 C. Yang and Y. Zhang, *J. Chem. Inf. Model.*, 2021, **61**, 4630–4644.
83 A. Gusach, A. Luginina, E. Marin, R. L. Brouillette, É. Besserer-Offroy, J.-M. Longpré, A. Ishchenko, P. Popov, N. Patel, T. Fujimoto, *et al.*, *Nat. Commun.*, 2019, **10**, 5573.
84 K. Kim, T. Che, O. Panova, J. F. DiBerto, J. Lyu, B. E. Krumm, D. Wacker, M. J. Robertson, A. B. Seven, D. E. Nichols, *et al.*, *Cell*, 2020, **182**, 1574–1588.
85 A. Meller, M. D. Ward, J. H. Borowsky, J. M. Lotthammer, M. Kshirsagar, F. Oviedo, J. L. Ferres and G. Bowman, *Biophys. J.*, 2023, **122**, 445a.

86 V. Scardino, J. I. Di Filippo and C. N. Cavasotto, *iScience*, 2023, **26**, 105920.
87 M. Holcomb, Y.-T. Chang, D. S. Goodsell and S. Forli, *Protein Sci.*, 2023, **32**, e4530.
88 I. Loshchilov and F. Hutter, *arXiv*, 2017, preprint, arXiv:1711.05101, DOI: **10.48550/arXiv.1711.05101**.
89 D. P. Kingma and J. Ba, *arXiv*, 2014, preprint, arXiv:1412.6980, DOI: **10.48550/arXiv.1412.6980**.
90 J. Zhu and Z. Gu, *CD Crossdock Benchmark Set for DiffBindFR*, 2024, DOI: **10.5281/zenodo.10816044**.