Supplementary Information

Knowledge-Guided Diffusion Model for 3D Ligand-Pharmacophore Mapping

Jun-Lin Yu[1], Cong Zhou[1], Xiang-Li Ning[1], Jun-Mou[1], Fan-Bo Meng[1], Jing-Wei Wu[1], Yi-Ting Chen[1], Biao-Dan Tang[1], Xiang-Gen Liu[2,*], and Guo-Bo Li[1,*]

[1]Key Laboratory of Drug Targeting and Drug Delivery System of Ministry of Education, Department of Medicinal Chemistry, West China School of Pharmacy, Sichuan University, Chengdu, Sichuan 610041, China

[2]College of Computer Science, Sichuan University, Chengdu 610065, China

*Correspondence: liguobo@scu.edu.cn (G.B.L.) or liuxianggen@scu.edu.cn (X.G.L.).

Contents

Supplementary Methods

SM. 1 DiffPhore architecture

DiffPhore consists of the following three main modules, including a knowledge-guided LPM representation encoder ($LPMEncoder$), a conformation generator ($CFGenerator$), and a calibrated conformation sampler ($CCSampler$). We here detail the architecture, implementation, and training process of DiffPhore.

(1) Knowledge-guided LPM representations

We proposed a heterogenous geometric graph G_t to characterize LPMs in the 3D space, serving as the input of the conformation generator,

$$G_t = LPMEncoder\left(\hat{G}_{l,t}, G_p\right) = \left\{\hat{G}_{l,t}, G_p, G_{lp}\right\} \quad (S1)$$

where $\hat{G}_{l,t} = \{\mathcal{V}_l, \mathbf{x}_t, \mathcal{E}_l, \mathbf{V}_l\}$ is a ligand graph, $G_p = \left\{\mathcal{V}_p, \mathbf{x_p}, \mathcal{E}_p, \mathbf{V}_p\right\}$ is a pharmacophore graph, and G_{lp} is a bipartite graph (Fig. 2b).

Here, we detail the ligand and pharmacophore graph representations. Ligand node features $\mathcal{V}_l$ include standard molecular descriptors, such as atomic number, degree, chirality, formal charge, connected hydrogens, implicit valence, radical electrons, hybridization, aromaticity, and ring membership (specifying sizes from 3 to 8). To help the model better understand the steric clash constraints, two dynamic boundary features are taken into consideration, which include the minimum distance d_{min} between the current node (ligand atom) and the nearest exclusion sphere, along with a range feature f_{range}, denoting whether the distance is in the range of a set of cutoffs ([1, 2, 3, 4, 5] Å). The dynamic boundary features are fused into the original ligand node features *via* concatenation. Ligand edge features consist of bond type (f_{bond}) and distance.

Each of the pharmacophore point $\mathcal{V}_p$ is featured by its type f_P, radius, whether or not it has normal direction, whether or not it is exclusion sphere, and especially the normal direction $\mathbf{f_d} = \left\{\mathbf{f_{d_{j'}}} \middle| j' \in \mathcal{V}_p\right\}$ if available. Edges in G_p are characterized by the distance between the connected nodes.

Most importantly, the bipartite graph $G_{lp} = \{\mathcal{V}_l, \mathcal{V}_p, \mathcal{E}_{lp}, \mathbf{V}_{lp}, \mathbf{N}_{lp}\}$ is exploited to describe the ligand-pharmacophore matching relations, where $\mathcal{E}_{lp}$ connects each ligand atom to all the pharmacophore feature points, $\mathbf{V}_{lp} = \{\mathbf{v}_{ij'} | \, i \in \mathcal{V}_l, j' \in \mathcal{V}_p\}$ stand for the pharmacophore type matching vectors, $\mathbf{N}_{lp} = \{\mathbf{n}_{ij'} | \, i \in \mathcal{V}_l, j' \in \mathcal{V}_p\}$ are the direction matching vectors. According to the definition of LPM rules in our AncPhore tool[1], we further included the following ligand node features to assist the construction of G_{lp}: (1) ligand atom pharmacophore fingerprints f_T, representing the possible pharmacophore types for each atom; (2) ligand atom pharmacophore orientations f_N, the vector pointing from the root to the considered atom for HA, HD, MB, and XB, or perpendicular to the aromatic ring plane for AR; (3) reference angles θ_{ref}, representing the angle between f_N and the ideal pharmacophore matching direction. The pharmacophore fingerprint feature (f_T) for each atom is represented as a vector defined by the potential of each pharmacophore type based on a predefined set of functional groups. This representation defines the potential of the current atom to fit with possible specific pharmacophore feature types.

The pharmacophore type matching vectors $\mathbf{V}_{lp} = \{\mathbf{v}_{ij'}\}$ are characterized by a set of weighted vectors between ligand atoms and pharmacophore points, which are utilized to update the node features.

$$\mathbf{v}_{ij'} = \left(\mathbf{y}_{j'} - \mathbf{x}_{t,i}\right) \cdot W_{match_{ij'}} \quad (S2)$$

$$W_{match_{ij'}} = Softmax_i\left(f_{match_{ij'}}\right) \quad (S3)$$

$$f_{match_{ij'}} = \varphi\left(f_{T_i},\ f_{Pj'},\ f_{T_{ij'}},\ d_{ij'}\right) \quad (S4)$$

$$f_{T_{ij'}} = f_{Ti} * f_{Pj'} \quad (S5)$$

The weights $W_{match_{ij'}}$ are calculated from the pharmacophore fingerprint feature f_{T_i} of the atom i, pharmacophore type $f_{Pj'}$ of the pharmacophore feature point j', the pharmacophore type matching feature $f_{T_{ij'}}$, and distance $d_{ij'}$ between them. φ refers to a MLP layer.

$$\mathbf{n}_{ij'} = \mathbf{v}_{ax_{ij'}} * \Delta\theta \quad (S6)$$

$$\mathbf{v}_{ax_{ij'}} = \frac{\left(\mathbf{f}_{N_i} * f_{T_{ij'}}\right) \times \mathbf{f}_{\boldsymbol{d}_{j'}}}{\left|\left(\mathbf{f}_{N_i} * f_{T_{ij'}}\right) \times \mathbf{f}_{\boldsymbol{d}_{j'}}\right|} \quad (S7)$$

$$\Delta\theta = \theta - \theta_{ref_i} * f_{T_{ij'}} \quad (S8)$$

The discrepancy between the ligand atom orientation and corresponding pharmacophore normal direction is described as a rotation transformation in the form of axis-angle vectors denoted as $\mathbf{N}_{lp} = \{\mathbf{n}_{ij'}\}$. The edge-specific atom orientation and ideal reference angle are determined by multiplying the pharmacophore type matching feature $f_{T_{ij'}}$; this product is then used to calculate the axis vector $\mathbf{v}_{ax_{ij'}}$ and the angle $\Delta\theta$. These approaches facilitate specific updates of the node features, adhering to the pharmacophore matching principles.

(2) Diffusion-based conformation generator

Given the random diffusion time t and LPM representations G_t, the

conformation generator is tasked with predicting the change directions (or scores) in the ligand translation ($\boldsymbol{\alpha}$), rotation ($\boldsymbol{\beta}$), and torsion angles ($\boldsymbol{\gamma}$).

$$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} = CFGenerator(G_t, t) \quad (S9)$$

The SE(3)-equivariant conformation generator (Supplementary Fig. 3) comprises the embedding, update, and output modules. The update module consists of L Message Passing layers, each with intra- and inter-graph update layers. The intra-graph layer extracts the topological features of the ligand and the pharmacophore separately; the inter-graph layer performs the feature fusion between two graphs, establishing deep representations of the ligand-pharmacophore interactions. Finally, the output module predicts the change directions in the ligand translation ($\boldsymbol{\alpha}$), rotation ($\boldsymbol{\beta}$), and torsion angles ($\boldsymbol{\gamma}$).

Initially, the embedding module processes input ligand and pharmacophore graphs (including numerical and categorical features and cross-edges between the graphs) and integrates the random diffusion time (t) into the graph features. This module contains two main parts: node embedding and edge embedding. To calculate the initial node embeddings (h_l^0, h_p^0), the categorical ($h_{l,cat},\ h_{p,cat}$) and numerical ($h_{l,num},\ h_{p,num}$) features are processed with Embedding and Linear layers from PyTorch package respectively as shown in Supplementary Fig. 3b and Eq. S10-S13. The diffusion time t is leveraged to calculate the sigma term according to the predefined noise schedule ($\sigma(t)$), which is injected into the node features.

$$h_{l,cat} = EmbeddingLayer\big(CONCAT([f_{l,cat}, f_{range}])\big) \quad (S10)$$

$$h_{l,num} = Linear\big(CONCAT([f_{l,num}, d_{min}, \sigma(t)])\big) \quad (S11)$$

$$h_{p,cat} = EmbeddingLayer\big(f_{p,cat}\big) \quad (S12)$$

$$h_{p,num} = Linear\big(CONCAT([f_{p,num}, \sigma(t)])\big) \quad (S13)$$

$$h_l^0 = h_{l,cat} + h_{l,num} \quad (S14)$$

$$h_p^0 = h_{p,cat} + h_{p,num} \quad (S15)$$

As for the edge features, radical bias embeddings of the edge length ($\mu(\cdot)$, calculated with a 'GaussianSmearing' layer, see source code) and the sigma term are concatenated and go through a Linear layer to obtain the initial edge embeddings (e_l, e_p, e_{lp}) as follows:

$$e_l = Linear\big(CONCAT([Distance(\mathcal{E}_l), \sigma(t), f_{bond}])\big) \quad (S16)$$

$$Distance(\mathcal{E}_l) = \{\mu(\|\boldsymbol{x}_{t,i} - \boldsymbol{x}_{t,j}\|) \mid i,j \in \mathcal{E}_l\} \quad (S17)$$

$$e_p = Linear\left(CONCAT\big([Distance(\mathcal{E}_p), \sigma(t)]\big)\right) \quad (S18)$$

$$Distance(\mathcal{E}_p) = \{\mu(\|\mathbf{x}_{p,i'} - \mathbf{x}_{p,j'}\|) \mid i',j' \in \mathcal{E}_p\} \quad (S19)$$

$$e_{lp} = Linear\left(CONCAT\big([Distance(\mathcal{E}_{lp}), \sigma(t)]\big)\right) \quad (S20)$$

$$Distance(\mathcal{E}_{lp}) = \{\mu(\|\mathbf{x}_{t,i} - \mathbf{x}_{p,j'}\|) \mid i,j' \in \mathcal{E}_{lp}\} \quad (S21)$$

Next, the update module iteratively refines the initial embeddings *via* message passing layers, with each layer comprising intra-graph and inter-graph updates. In detail, Intra-graph updates compute messages ($m_{l,i-intra}$ and $m_{p,j'-intra}$, Eq. S22-S23) as tensor products of the node features and the spherical harmonic representations of neighboring edge vectors, weighted by the edge embedding e_{ij}, the outgoing h_j and the incoming node features h_i.

$$m_{l,i-intra} = TP_{l\to l}\big(h_{l,i}^{l-1}, \mathbf{V_l}, e_l\big) \quad (S22)$$

$$m_{p,j'-intra} = TP_{p\to p}\big(h_{p,j'}^{l-1}, \mathbf{V_p}, e_p\big) \quad (S23)$$

$$TP(h_i, \mathbf{V}, e) = BN\left(\frac{1}{|\mathcal{N}_i|}\sum_{j\in\mathcal{N}_i, \mathbf{v}_{ij}\in\mathbf{V}} Y(\mathbf{v}_{ij}) \otimes \phi(e_{ij}, h_i, h_j)\, h_j\right) \quad (S24)$$

In the formulas, TP denotes the tensor products layer, $\mathcal{N}_i = \{j|(i,j) \in \mathcal{E}_l \; or \; \mathcal{E}_p\}$ stands for neighbor nodes and Y is the spherical harmonic. BN refers to the batch normalization, ϕ stands for a MLP layer, and $\otimes$ refers to tensor product operation. These tensor product layers are implemented using the 'FullyConnectedTensorProduct' layer from the E3NN[2] package. The process leverages either related positions or production kernels, resulting in SE(3)-equivariant representations that are invariant to rotation and translation and providing unbiased 3D representations of ligand-pharmacophore interactions.

The inter-graph layer simulates the ligand-pharmacophore recognition and alignment with the constructed bipartite graph G_{lp}. The pharmacophore type and direction matching vectors are incorporated for pharmacophore fitting, by using similar tensor product layers (Eq. S25-S26). The intra- and inter-graph messages are aggregated to update ligand (h_l^l) and pharmacophore (h_p^l) node embeddings (Eq. S27-S28).

$$m_{l,i-inter} = TP_{p\to l,type}\left(h_{l,i}^{l-1}, \mathbf{V}_{lp}, e_{lp}\right) + TP_{p\to l,direction}\left(h_{l,i}^{l-1}, \mathbf{N}_{lp}, e_{lp}\right) \quad (S25)$$

$$m_{p,j'-inter} = TP_{l\to p,type}\left(h_{p,j'}^{l-1}, \mathbf{V}_{lp}, e_{lp}\right) + TP_{l\to p,direction}\left(h_{p,j'}^{l-1}, \mathbf{N}_{lp}, e_{lp}\right) \quad (S26)$$

$$h_{l,i}^{l} = h_{l,i}^{l-1} + m_{l,i-intra} + m_{l,i-inter} \quad (S27)$$

$$h_{p,j'}^{l} = h_{p,j'}^{l-1} + m_{p,j'-intra} + m_{p,j'-inter} \quad (S28)$$

As the iteration progresses, the node features of both ligand and pharmacophore graphs are systematically adjusted according to pharmacophore principles, with the constraints of the pharmacophore model.

Finally, the output module receives the updated ligand features h_l^L to predict the translation, rotation and torsion scores *w.r.t* their diffusion kernels.

The translation and rotation are rigid transformations operating on the center of mass of the ligand so that a graph $G_{cl} = \{c, \mathcal{V}_l, \mathcal{E}_{cl}, \mathbf{V}_{cl}\}$ connecting ligand atoms $a \in \mathcal{V}_l$ and the center of mass c are constructed to compute the corresponding scores $\boldsymbol{\alpha}, \boldsymbol{\beta}$ with tensor product layer (Φ, implemented as the 'FullyConnectedTensorProduct' layer from E3NN package). The vectors $\mathbf{V}_{cl} = \{\mathbf{v}_{ca} | \, a \in \mathcal{V}_l\}$ stand for the 3D vectors from the center of mass to the ligand atom a. The edge features e_{ca} is calculated according to the edge length. The tensor product layer aggregates all the ligand node features to the ligand center and yield the scores $\boldsymbol{\alpha}, \boldsymbol{\beta}$.

$$e_{ca} = \psi\big(\mu(\|\mathbf{v}_{ca}\|)\big), a \in \mathcal{V}_l \qquad (S29)$$

$$\boldsymbol{\alpha}, \boldsymbol{\beta} = \Phi(G_{cl}) = \frac{1}{|\mathcal{V}_l|} \sum_{a \in \mathcal{V}_l, \mathbf{v}_{ca} \in \mathbf{V}_{cl}} Y(\mathbf{v}_{ca}) \otimes \phi\big(e_{ca}, h_{l,a}^L\big) \, h_{l,a}^L \qquad (S30)$$

Here, ψ, ϕ are MLP layers, and $\mu(\cdot)$ denotes radical bias embedding of edge length.

To estimate the torsion score $\boldsymbol{\gamma}$, we constructed another graph $G_{rl} = \{\mathcal{R}, \mathcal{V}_l, \mathcal{E}_{rl}, \mathbf{V}_{rl}, \mathbf{V}_r\}$ connecting the rotatable bonds ($r \in \mathcal{R}$) to their adjacent atoms ($b \in \mathcal{V}_l$). $\mathcal{E}_{rl}$ are the edges between them. $\mathbf{V}_{rl} = \{\mathbf{v}_{\mathrm{rb}}\}$ and $\mathbf{V}_r = \{\mathbf{v}_r\}$ are respectively the vectors from the bond center of r to the atom b and from the incoming node to the outgoing node of bond r. The embedding of the rotatable bonds (h_r) are calculated as the concatenation of the involved node features of the incoming node $h_{r_{in}}$ and the outgoing node $h_{r_{out}}$. As the rotatable bond r is undirected, the incoming and outgoing nodes can be exchanged. The edge features e_{rb} is obtained through a radical bias embedding of the distance from the bond center to the ligand atom, where ω refers to an MLP layer.

$$h_r = CONCAT\left(\left[h_{r_{in}}, h_{r_{out}}\right]\right) \quad (S31)$$

$$e_{rb} = \omega\left(\mu(\|\mathbf{v}_{rb}\|)\right) \quad (S32)$$

The torsion score $\boldsymbol{\gamma}$ is obtained by aggregating adjacent atom features to the rotatable bonds via a similar tensor product layer (φ, implemented as the 'FullTensorProduct' layer from E3NN package). For the torsion score $\boldsymbol{\gamma}_{\mathbf{r}}$ of the rotatable bond r, φ performs the tensor product operation between the bond features and the adjacent atom features, given by

$$\boldsymbol{\gamma} = \varphi(G_{rl}) \quad (S33)$$

$$\boldsymbol{\gamma}_{\mathbf{r}} = \frac{1}{|\mathcal{N}_r|} \sum_{b \in \mathcal{N}_r, \mathbf{v}_{rb} \in \mathbf{V}_{rl}} T_r(\mathbf{v}_{rb}) \otimes \phi(e_{rb}, h_r, h_b^L)\, h_b^L \quad (S34)$$

$$T_r \coloneqq Y^2(\mathbf{v}_r) \otimes Y(\mathbf{v}_{rb}) \quad (S35)$$

where $\mathcal{N}_r$ is the set of ligand atoms connected with the rotatable bond. T_r is a convolutional filter constructed for each rotatable bond r calculating the tensor product of the spherical harmonic representation of the bond axis $\mathbf{v}_r$ (Y^2 here means max level is 2) and the vector $\mathbf{v}_{rb}$ from bond center of r to the atom b. ϕ is an MLP layer.

(3) Calibrated conformation sampler

The calibrated conformation sampler (CCSampler) mimics the inference process[3,4] to construct pseudo ligand conformations and correspond scores (as labels) for model training as shown in the pseudo code in Supplementary Table 3. The pseudo ligand conformations $\tilde{\mathbf{x}}_t$ are estimated based on the denoised data points by DiffPhore based on the conformation in the last step $\mathbf{x}_{t+\Delta t}$, given by

$$\widehat{\boldsymbol{\alpha}}_{t+\Delta t}, \widehat{\boldsymbol{\beta}}_{t+\Delta t}, \hat{\boldsymbol{\gamma}}_{t+\Delta t} = CFGenerator\left(LPMEncoder\left(\hat{G}_{l,t}, G_p\right), t\right) \quad (S36)$$

$$\mathbf{\Delta r}, \mathbf{\Delta R}, \mathbf{\Delta \theta} = Estimate(\widehat{\boldsymbol{\alpha}}_{t+\Delta t}, \widehat{\boldsymbol{\beta}}_{t+\Delta t}, \hat{\boldsymbol{\gamma}}_{t+\Delta t}) \quad (S37)$$

$$\tilde{\mathbf{x}}_t = A\big((\mathbf{\Delta r}, \mathbf{\Delta R}, \mathbf{\Delta \theta}),\ \mathbf{x}_{t+\Delta t}\big) \quad (S38)$$

where the function $Estimate$ samples transformation $(\mathbf{\Delta r}, \mathbf{\Delta R}, \mathbf{\Delta \theta})$ according to the predicted scores $(\widehat{\boldsymbol{\alpha}}_{t+\Delta t}, \widehat{\boldsymbol{\beta}}_{t+\Delta t}, \hat{\boldsymbol{\gamma}}_{t+\Delta t})$ and the predefined variances of the diffusion kernels[5] (see the pseudo codes in Supplementary Table 3). Since the transformation of the conformation at the last step $\mathbf{\Delta r}, \mathbf{\Delta R}, \mathbf{\Delta \theta}$ compared to the original conformation is known, the transformation $\mathbf{\Delta r}_{0\to t}, \mathbf{\Delta R}_{0\to t}, \mathbf{\Delta \theta}_{0\to t}$ at step *t* corresponding to the above pseudo conformation can be directly calculated, given by the following formula:

$$\boldsymbol{\Delta \theta}_{0\to t} = \boldsymbol{\Delta \theta}_{0\to t+\Delta t} + \boldsymbol{\Delta \theta} \quad (S39)$$

$$\mathbf{x}_t^{tor} = A_{tor}(\mathbf{x}_0,\ \boldsymbol{\Delta \theta}_{\mathbf{0}\to \boldsymbol{t}}) \quad (S40)$$

$$\mathbf{R},\ \mathbf{T} = Superimpose(\mathbf{x}_t^{tor},\ \tilde{\mathbf{x}}_t) \quad (S41)$$

$$\boldsymbol{\Delta} \mathbf{r}_{0\to t} = \mathbf{T} + mean(\mathbf{x}_t^{tor})@\mathbf{R} - mean(\mathbf{x}_t^{tor}) \quad (S42)$$

$$\boldsymbol{\Delta} \mathbf{R}_{0\to t} = \mathbf{R} \quad (S43)$$

where the function $Superimpose$ refers to the rigid alignment *via* the Kabsch–Umeyama algorithm[6], which is used to estimate the rotation angles in the Hilbert space. Based on the above transformation corresponding to the pseudo conformation $\tilde{\mathbf{x}}_t$, the scores can be computed according to following gradient function[5].

$$\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\gamma}}_{\boldsymbol{t}} \leftarrow \nabla p_t^{tr}(\boldsymbol{\Delta} \mathbf{r}_{\mathbf{0}\to \boldsymbol{t}}|0),\ \nabla p_t^{rot}(\boldsymbol{\Delta} \mathbf{R}_{\mathbf{0}\to \boldsymbol{t}}|0)\ ,\ \nabla p_t^{tor}(\boldsymbol{\Delta \theta}_{\mathbf{0}\to \boldsymbol{t}}|0) \quad (S44)$$

In this way, the pseudo conformation $\tilde{\mathbf{x}}_t$ with its target labels $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\gamma}}_{\boldsymbol{t}}$ at step t form a training data point for DiffPhore, which reserves the most information of the real data point and also introduces the estimation error in the inference process. The detailed implementation and computations of the

process can be found in the source codes.

(4) Training details

The model is firstly warmed-up on the well-designed LigPhoreSet to learn the essential rules of ligand conformation mapping with pharmacophore features. At the initial training stages, we used the 10% subset of LigPhoreSet to search hyperparameters. Then, the model was trained on the entire LigPhoreSet dataset, using a split ratio of 8:1:1 for the training set, validation set, and test set. We then refined the model with the CpxPhoreSet to help it get a deeper sight into the real-world LPM scenarios with also the time-split dataset division. Notably, the CCSampler strategy was adopted in the refinement stage.

DiffPhore is implemented with PyTorch, PyTorch-Geometric and E3NN library. We trained it with NVIDIA RTX 4090 GPU and Intel(R) Xeon(R) Platinum 8378C CPU @ 2.80GHz. The final DiffPhore model has 2.23 million parameters and is trained for 40 and 800 epochs respectively in the warm-up stage (12 days) and the refining stage (10 days). Due to the high computational cost of inference, the model is validated by calculating the inference success rate and fitness (*DfScore1*) every 5 epochs (denoising step = 20, number of initial poses = 1), using at most 1000 complex structures from the validation set. The validation success rate and fitness score are then used to select the best model variant. The batch size is set 10 because of memory limitation and learning rate is 1e-3. The detailed hyperparameters are provided in the Supplementary Table 11 and source code repository.

SM. 2 Evaluation metrics

There are several key metrics involved in the performance evaluations,

including Root Mean Square Deviation (RMSD), PoseBusters test validity (PB-Valid), Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC), Enrichment Factor (EF) and Area Under the Receiver Operating Characteristic Curve (AUROC).

RMSD quantifies the average spatial discrepancy between the predicted atomic coordinates $\{\boldsymbol{x}_i\}$ and the ground truth positions $\{\hat{\boldsymbol{x}}_i\}$:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2} \qquad (S45)$$

The PoseBusters test suite[7] is organized into three groups of tests: (1) chemical validity and consistency; (2) intramolecular validity; (3) intermolecular validity. '%RMSD<2Å & PB-Valid' denotes the percentage of predicted poses that pass all PoseBusters validity tests and exhibit an RMSD of less than 2 Å compared to the ground truth poses. '%RMSD<2Å & PB-Valid (without protein)' denotes the percentage of predicted poses that pass all PoseBuster validity tests, excluding the 'intermolecular validity' test with protein, while also having an RMSD of less than 2 Å relative to the ground truth poses.

BEDROC emphasizes the early identification of active compounds (Eq. S46), where n refers to the number of active ligands, N is the total number of ligands, r_i is the rank of the ith active ligand, R_a is the percentage of active ligands, α is a hyperparameter (set as 80.5).

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-\frac{\alpha r_i}{N}}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\frac{\alpha}{N}}-1}\right)}\left(\frac{R_a\, sinh\left(\frac{\alpha}{2}\right)}{cosh\left(\frac{\alpha}{2}\right)-cosh\left(\frac{\alpha}{2}-\alpha R_a\right)}\right)+\frac{1}{1-e^{\alpha(1-R_a)}} \qquad (S46)$$

EF_α quantifies the proportion of active ligands found within the top-ranked predictions, reflecting the model's ability to prioritize actives.

$$EF_\alpha = \frac{n_{top\alpha}/N_{top\alpha}}{n/N} \quad (S47)$$

AUROC comprehensively assesses the model's ability to discriminate between active and decoy compounds across a range of thresholds.

SM. 3 sQC/gQC/PGP-1 protein expression and purification

We followed the protocols from our previous study[8] for the expression and purification of human sQC (amino acids 33-361), gQC (amino acids 53-382), and the auxiliary enzyme PGP-1 (amino acids 1-215)[8]. In brief, recombinant plasmids encoding N-terminal Trx-His_6-TEV tags were cloned into the pET32a vector and transformed into *Escherichia coli* BL21-CodonPlus (DE3)-RIL cells. Protein expression was induced with 0.3 mM IPTG, followed by incubation at 16 °C for 48 hours. Cells were harvested by centrifugation at 1,753 xg for 20 minutes, resuspended in lysis buffer (50 mM Na_3PO_4, 300 mM NaCl, pH 8.0), and protease inhibitors, followed by cell lysis using an ultrahigh-pressure homogenizer. The resulting lysate was centrifuged at 15,777 xg for 30 minutes to obtain the supernatant.

Nickel-ion affinity chromatography-based protein purification was conducted using a pre-equilibrated Cube Biotech column with buffer A (50 mM Tris-HCl, 150 mM NaCl, pH 8.0, for sQC; 50 mM Tris-HCl, 150 mM NaCl, 5% glycerol, pH 7.5, for gQC). Target proteins were eluted in buffer B (50 mM Tris-HCl, 150 mM NaCl, 300 mM imidazole, pH 8.0, for sQC; 50 mM Tris-HCl, 150 mM NaCl, 5% glycerol, 300 mM imidazole, pH 7.5, for gQC). Fractions

containing sQC/gQC were concentrated by centrifugation and exchanged with imidazole-free buffer A using a 30 kDa cutoff Amicon Ultra (Millipore) prior to storage at −80 °C for activity assays.

TEV protease cleavage was performed overnight at 4 °C to remove the Trx-His_6 tag from protein samples. Further purification was conducted using Ni-NTA column. The final protein purity (>95%) was confirmed by SDS-PAGE and Coomassie staining. Purified samples were flash-frozen in liquid nitrogen and stored at −80 °C for subsequent thermostability and crystallization studies.

A similar expression and purification protocol was employed for the auxiliary enzyme PGP-1 (amino acids 1-215). The PGP-1 gene was cloned into the pET28 vector with a tag of N-terminal His_6-TEV. Following cultivation, cells were resuspended in PGP-1 buffer A (50 mM Tris-HCl, 500 mM NaCl, pH 8.0), lysed, and subjected to Ni-NTA chromatography with elution in PGP-1 buffer B (50 mM Tris-HCl, 500 mM imidazole, 500 mM NaCl, pH 8.0). Eluted fractions were further processed for buffer exchange using PGP-1 buffer C (25 mM Tris-HCl, 50 mM NaCl, pH 8.0). Purified PGP-1 was stored at −80 °C for subsequent activity assays.

SM. 4 sQC/gQC/PGP-1 inhibition activity assays

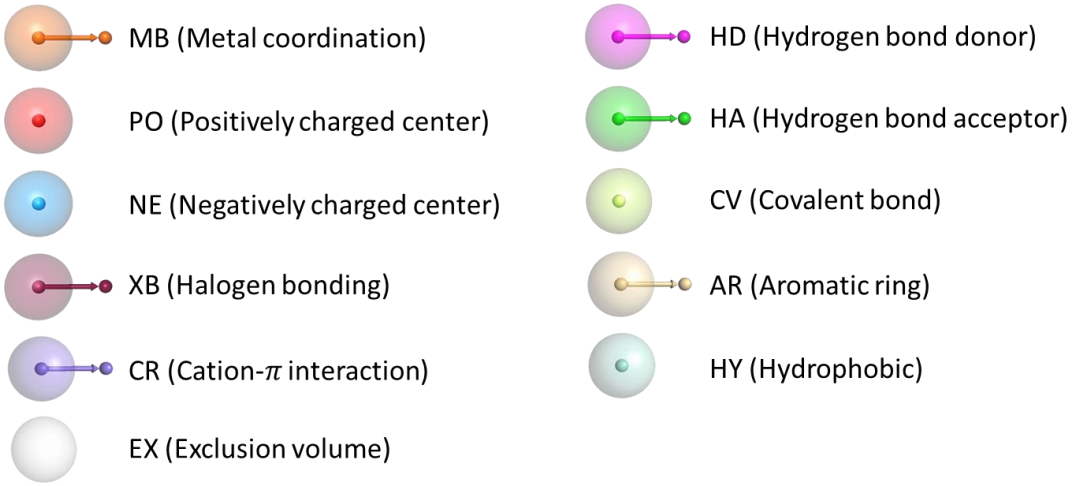
All compounds were tested for their inhibitory activity on sQC, gQC, and PGP-1 in the assay buffer (25 mM Tris-HCl, 150 mM NaCl, 10% glycerol, pH 8.0) as described previously[8]. Briefly, varying concentrations of compounds were firstly incubated with sQC (30 nM) or gQC (30 nM) for 10 minutes. Following this pre-incubation, PGP-1 (100 nM) and H-Gln-AMC (3 μM) were added to initiate the enzyme reactions. The fluorescence intensity was recorded using a microplate

reader with excitation and emission wavelengths set at 380 nm and 460 nm, respectively, over a 15-minute period. Similar to the sQC/gQC activity tests, the PGP-1 activity was evaluated for all compounds using pGA as the substrate. In these PGP-1 activity assays, a 60 μL reaction solution contained 50 nM PGP-1 and 1 μM pGA. Each experiment was conducted in triplicate, and the resulting data were analyzed using GraphPad Prism software to determine the IC_{50} values. The K_i values were obtained from the IC_{50} values using the Cheng−Prusoff equation[9]:

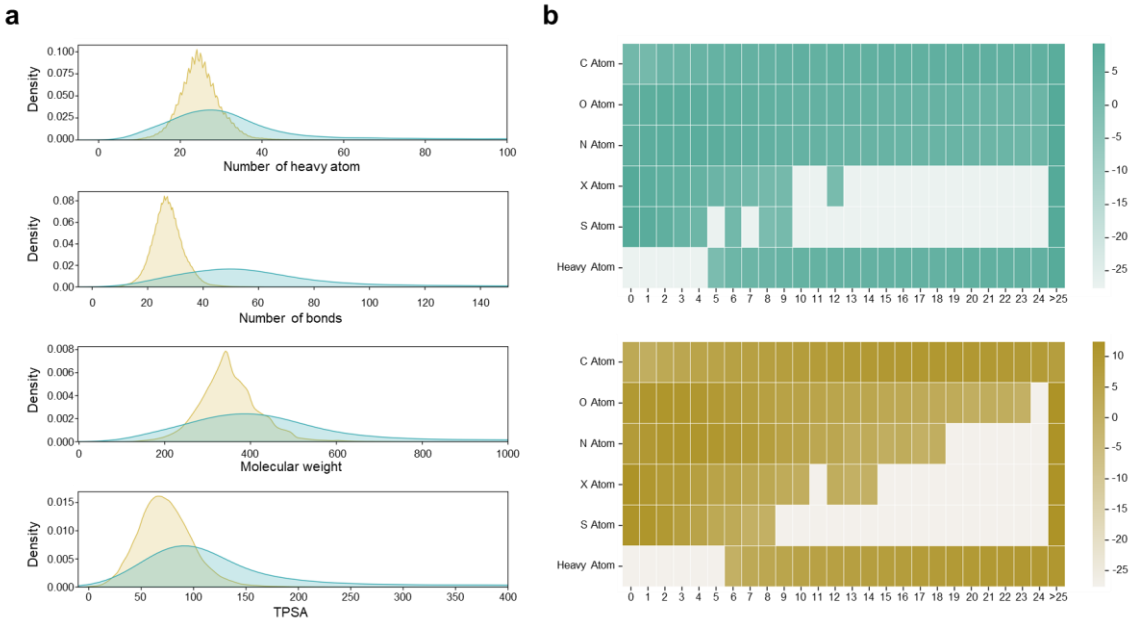
$$K_i = \frac{IC_{50}}{1 + \frac{[S]}{K_M}} \tag{S48}$$

where [S] is the H-Gln-AMC concentration for activity test and the Michaelis constant K_M values of sQC and gQC for cyclization of H-Gln-AMC are determined to be 87.77 ± 8.63 μM and 108.6 ± 13.63 μM, respectively.

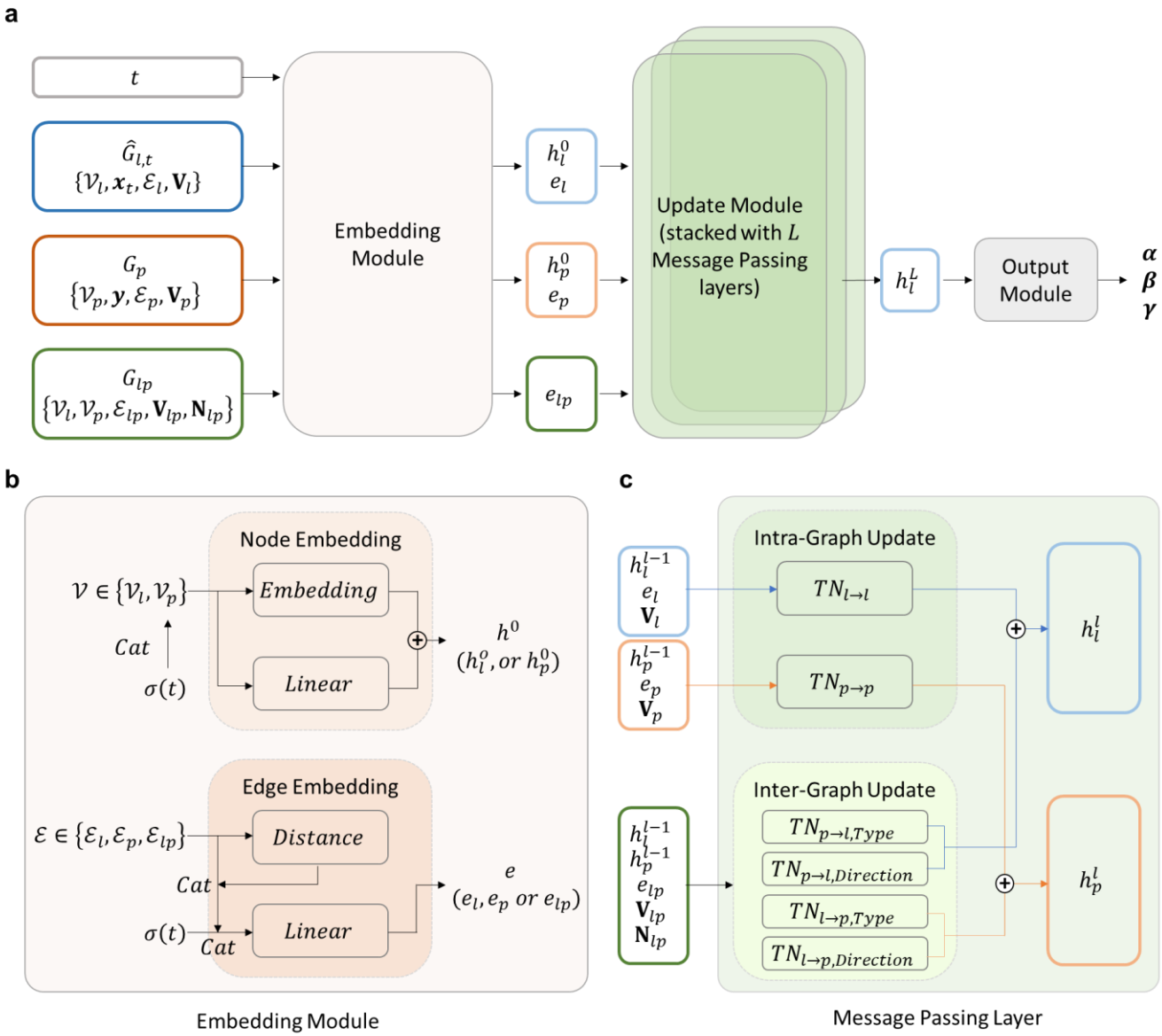
Supplementary Figures



Supplementary Fig. 1 The pharmacophore feature types and their corresponding color schemes.

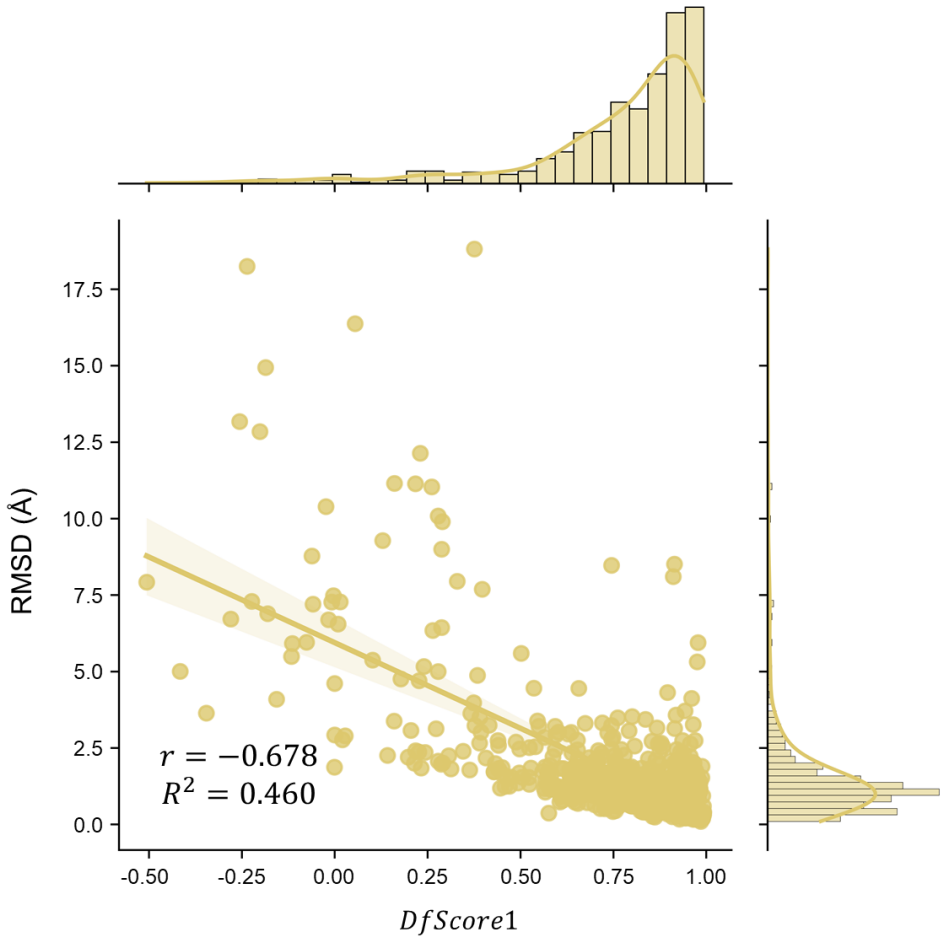


Supplementary Fig. 2 The comparison of ligand properties between CpxPhoreSet (cyan) and LigPhoreSet (yellow). **a** The distribution of number of heavy atoms, number of bonds, molecular weight and TPSA. **b** The heatmap of counts of different atoms (X means the total count of F, Cl, Br, and I atoms).

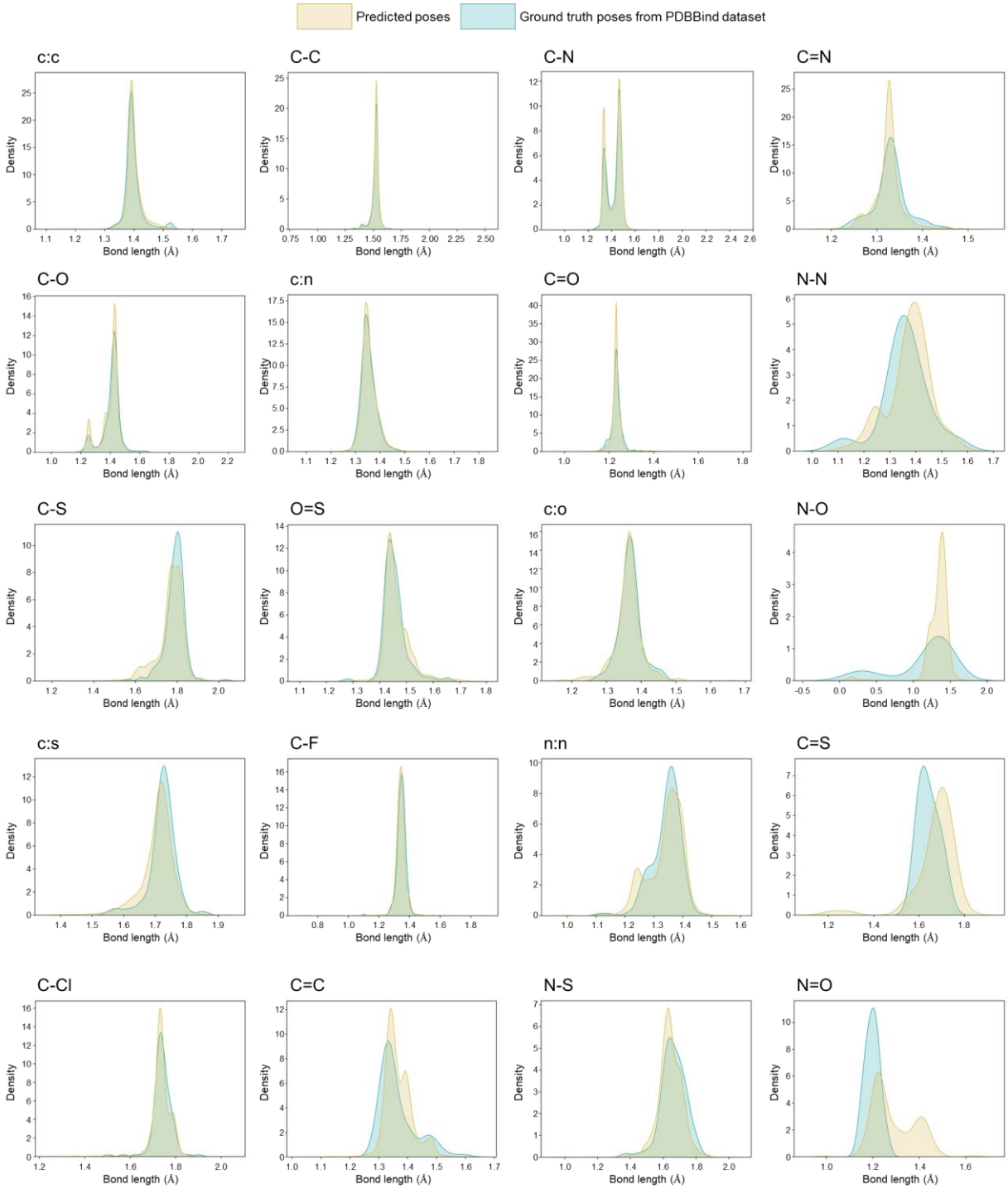


Supplementary Fig. 3 Architecture of the SE(3)-equivariant conformation generator. **a** Overall architecture. The generator takes the LPM representations $G_t = \{\hat{G}_{l,t}, G_p, G_{lp}\}$ and random diffusion time t as input, where $\hat{G}_{l,t}$ is the perturbed ligand graph, G_p is the pharmacophore graph, G_{lp} is the bipartite graph representing their interactions. The generator consists of three main components: Embedding, Update and Output modules. **b** Embedding module. This module processes the original node and edge features, incorporating diffusion time (t). In detail, the categorical and numerical features ($\mathcal{V}_l$ for ligand atoms, $\mathcal{V}_p$ for pharmacophore points) are respectively embedded using Embedding and Linear layers from PyTorch and summed to create the initial node embedding (h_l^0, h_p^0). Diffusion time (t) modulates the noise schedule ($\sigma(t)$) integrated into node and edge features. Edge distances (ligand edges $\mathcal{E}_l$, pharmacophore edges $\mathcal{E}_p$, and inter-graph edges $\mathcal{E}_{lp}$) are enhanced by a 'GaussianSmearing' module in '$Distance$' component that computes distance embeddings. The original edge features, concatenated with the distance

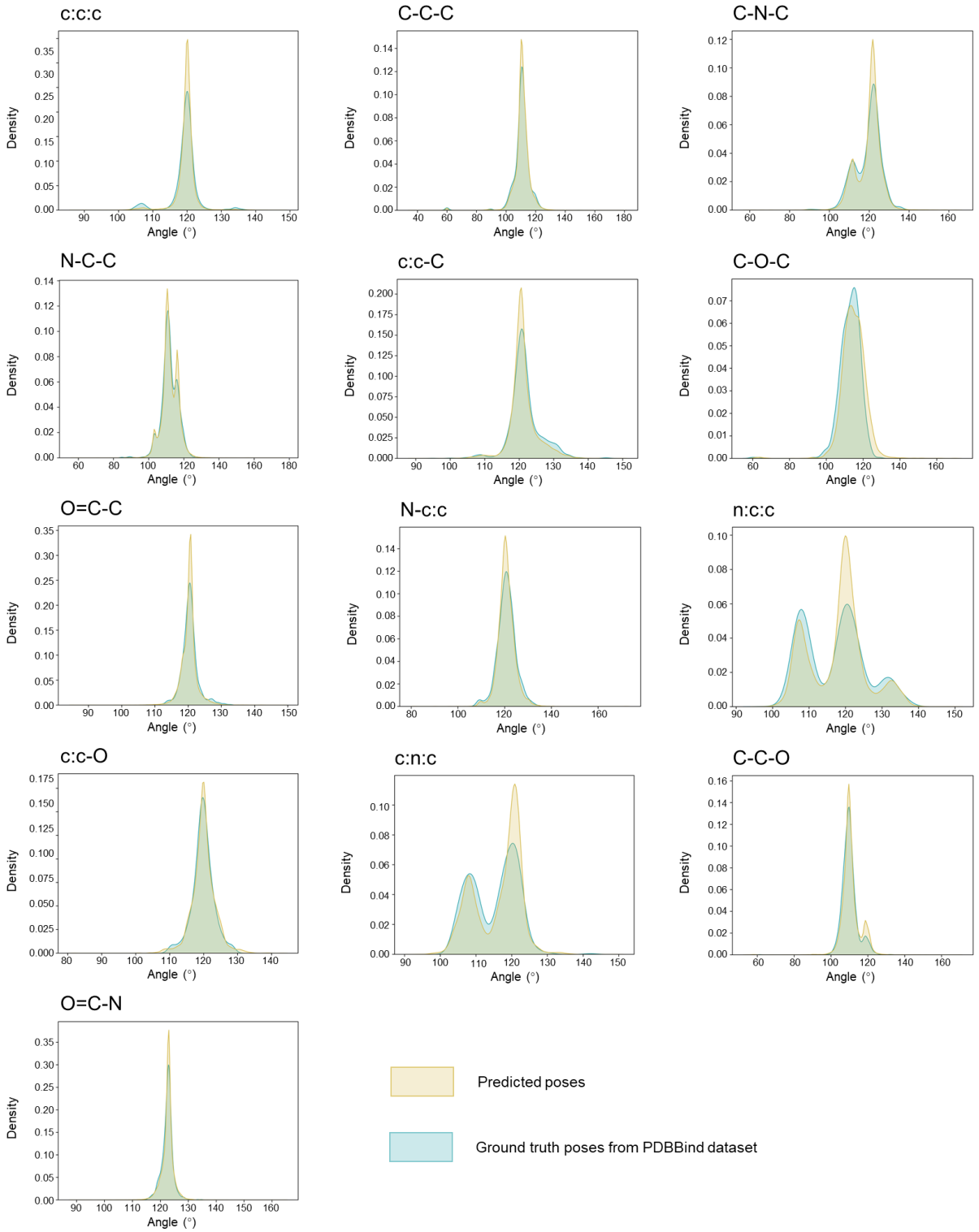
embedding as well as noise level ($\sigma(t)$), are then processed through a Linear layer to produce initial edge embeddings (e_l, e_p, e_{lp}). **c** Update module. This module is stacked with L Message Passing layers, with each computing the intra- and inter-graph updates of the input embeddings (h_l^{l-1}, h_l^{l-1}). The intra-graph updates are calculated as tensor products of the node features and the spherical harmonic representations of neighboring edge vectors ($\boldsymbol{V}_l$ or $\boldsymbol{V}_p$) within single graph, weighted by the edge embedding e_{ij}, the outgoing h_j and the incoming node features h_i. These tensor product operations are implemented using 'FullyConnectedTensorProduct' layer from the E3NN package. Inter-graph updates utilize the well-tailored pharmacophore type ($\boldsymbol{V}_{lp}$) and direction ($\boldsymbol{N}_{lp}$) matching vectors. All updates are aggregated to produce updated embeddings (h_l^{l-1}, h_l^{l-1}).



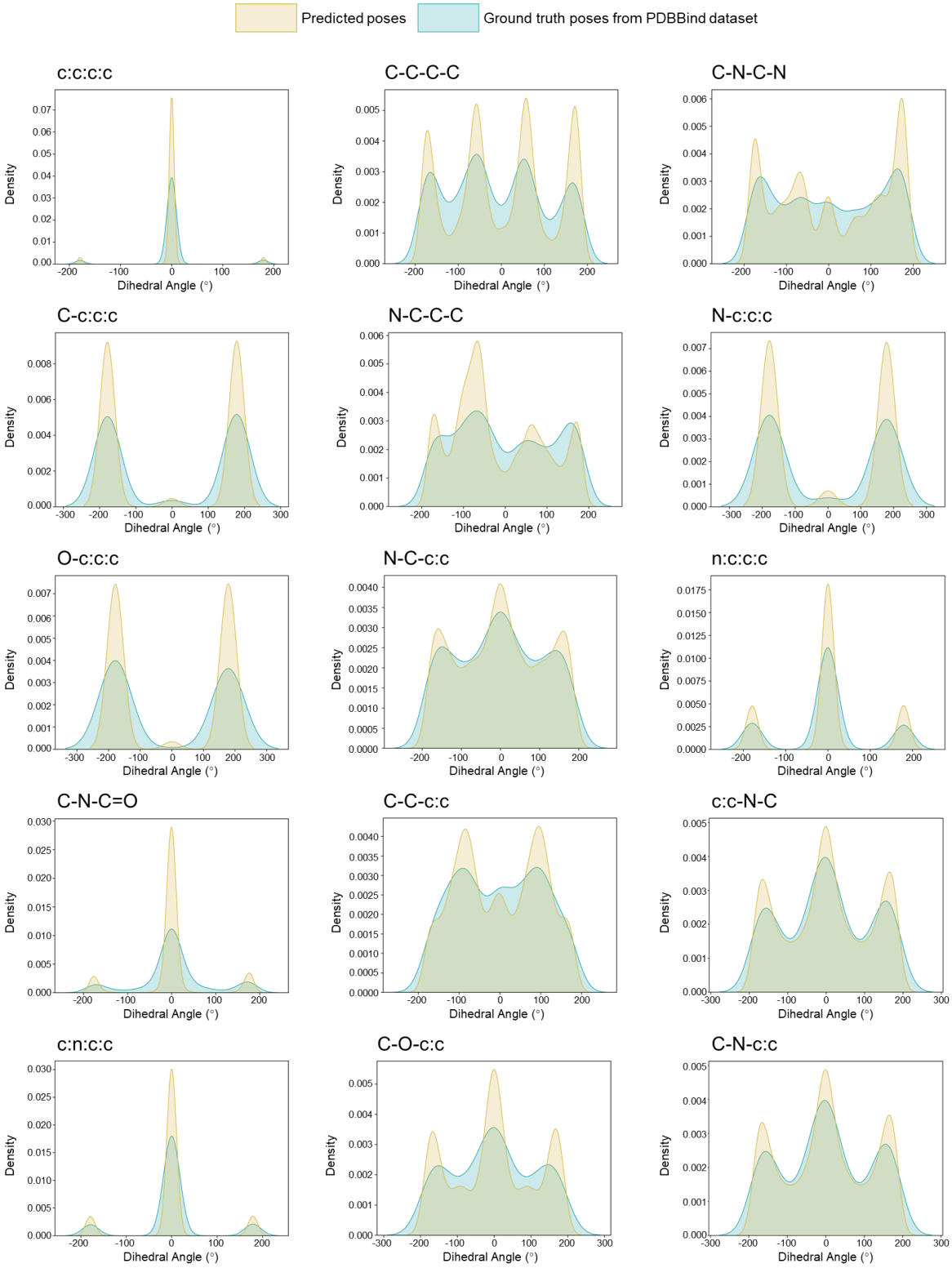
Supplementary Fig. 4 The correlations between *DfScore*1 and the RMSD values of all the predicted poses of DiffPhore on the PDBBind test set and PoseBusters set.



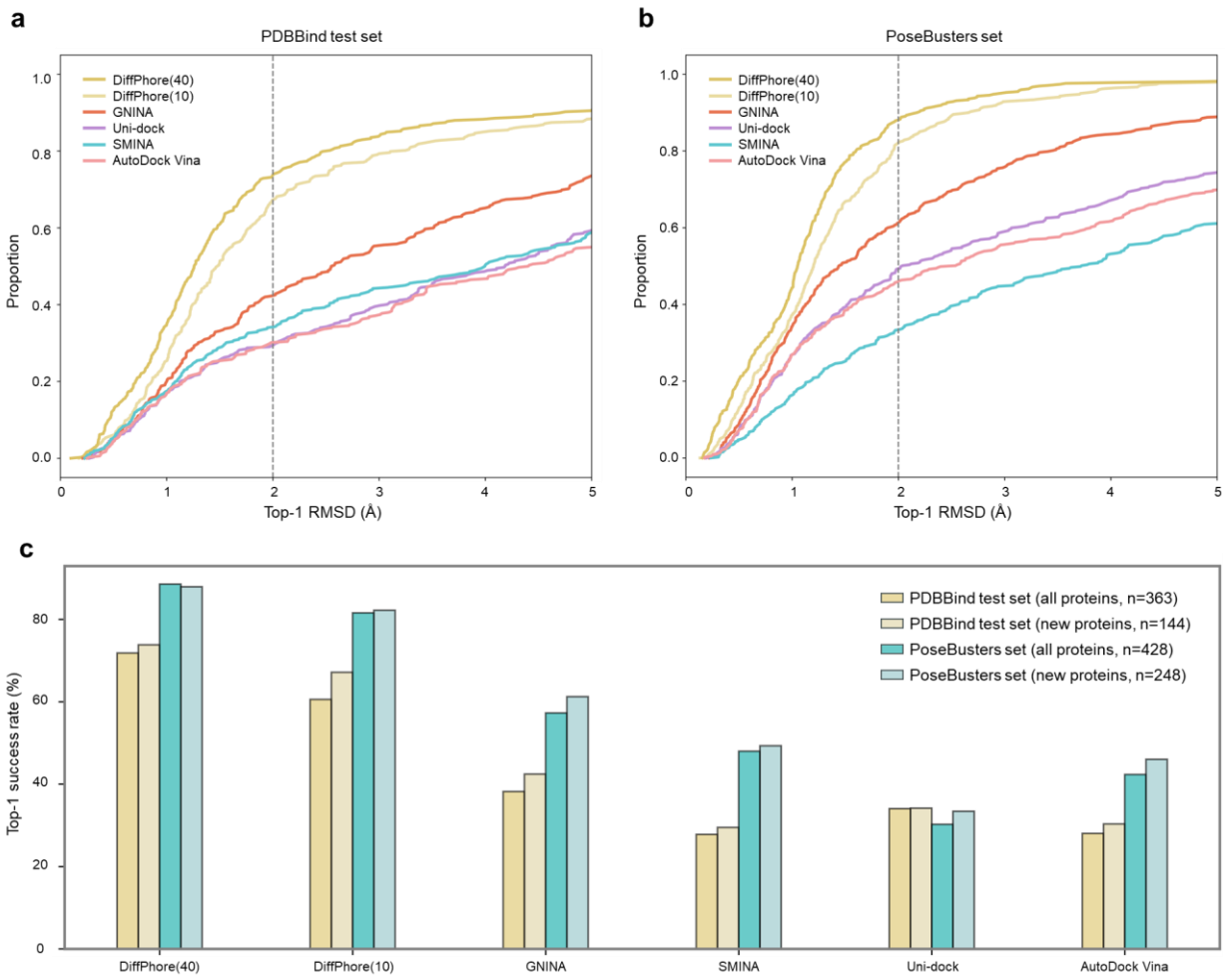
Supplementary Fig. 5 Comparison of bond lengths in the predicted conformations with the ground truth conformations from the PDBBind dataset. The element in lower case means the atom is in an aromatic ring; ':' refers to aromatic bond.



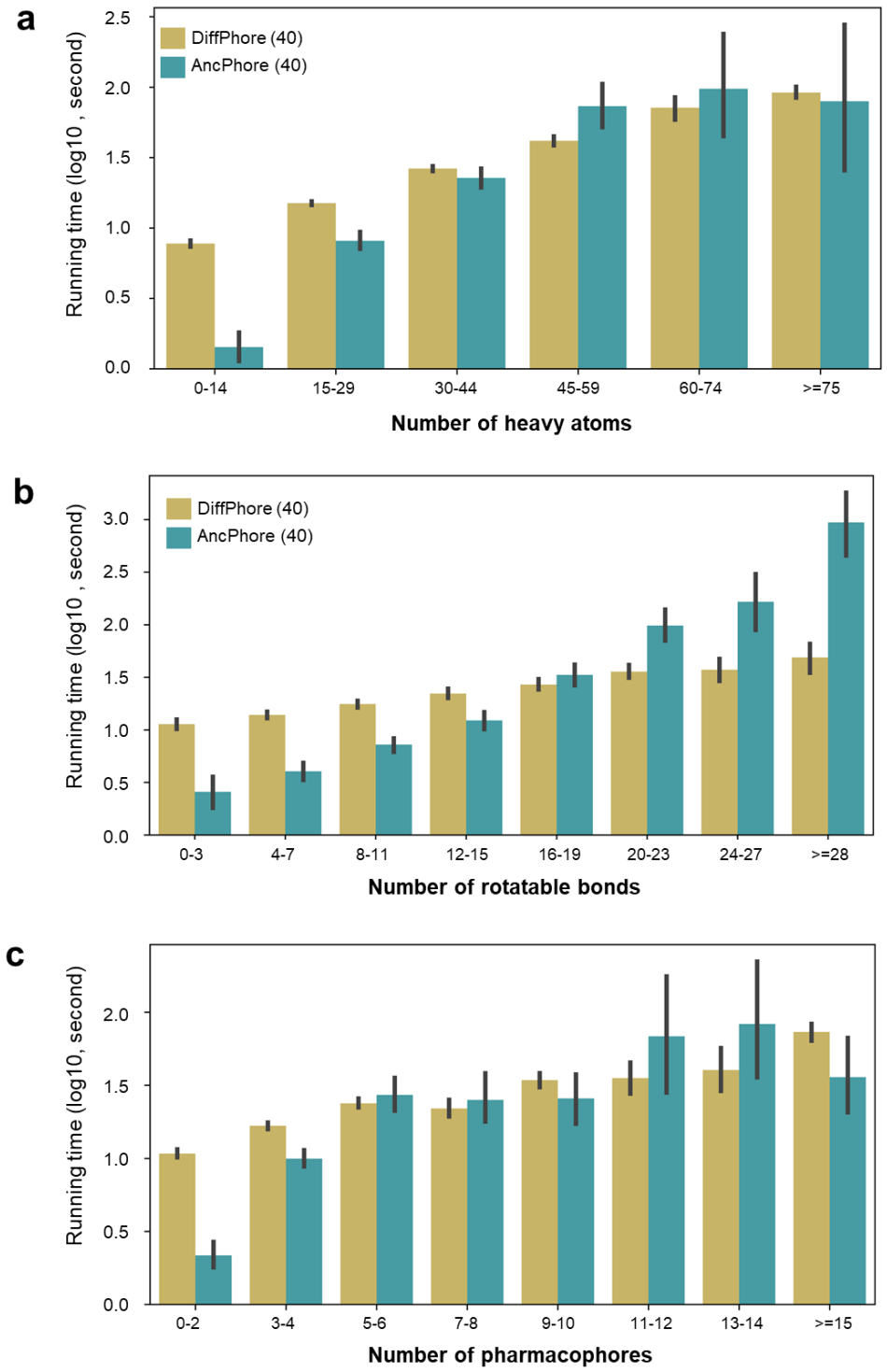
Supplementary Fig. 6 Comparison of angles in the predicted conformations with the ground truth conformations from the PDBBind dataset. The element in lower case means the atom is in an aromatic ring; ':' refers to aromatic bond.



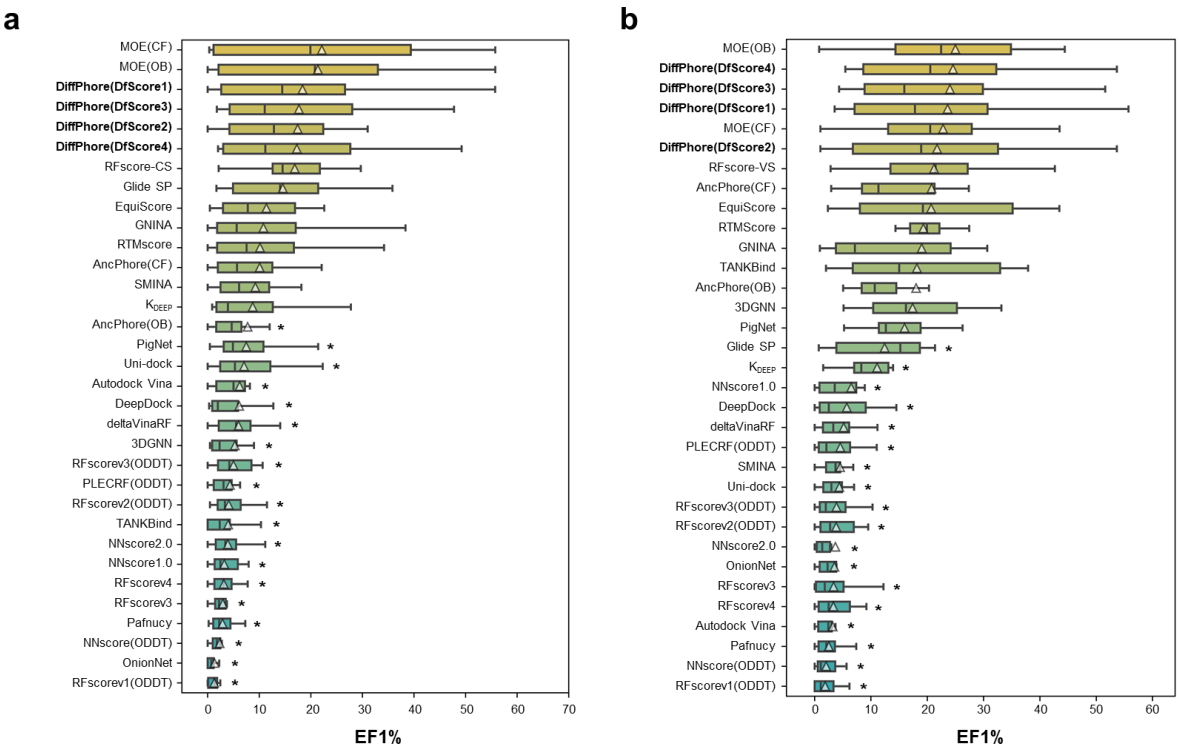
Supplementary Fig. 7 Comparison of dihedral angles in the predicted conformations with the ground truth poses from the PDBBind dataset. The element in lower case means the atom is in an aromatic ring; ':' refers to aromatic bond.



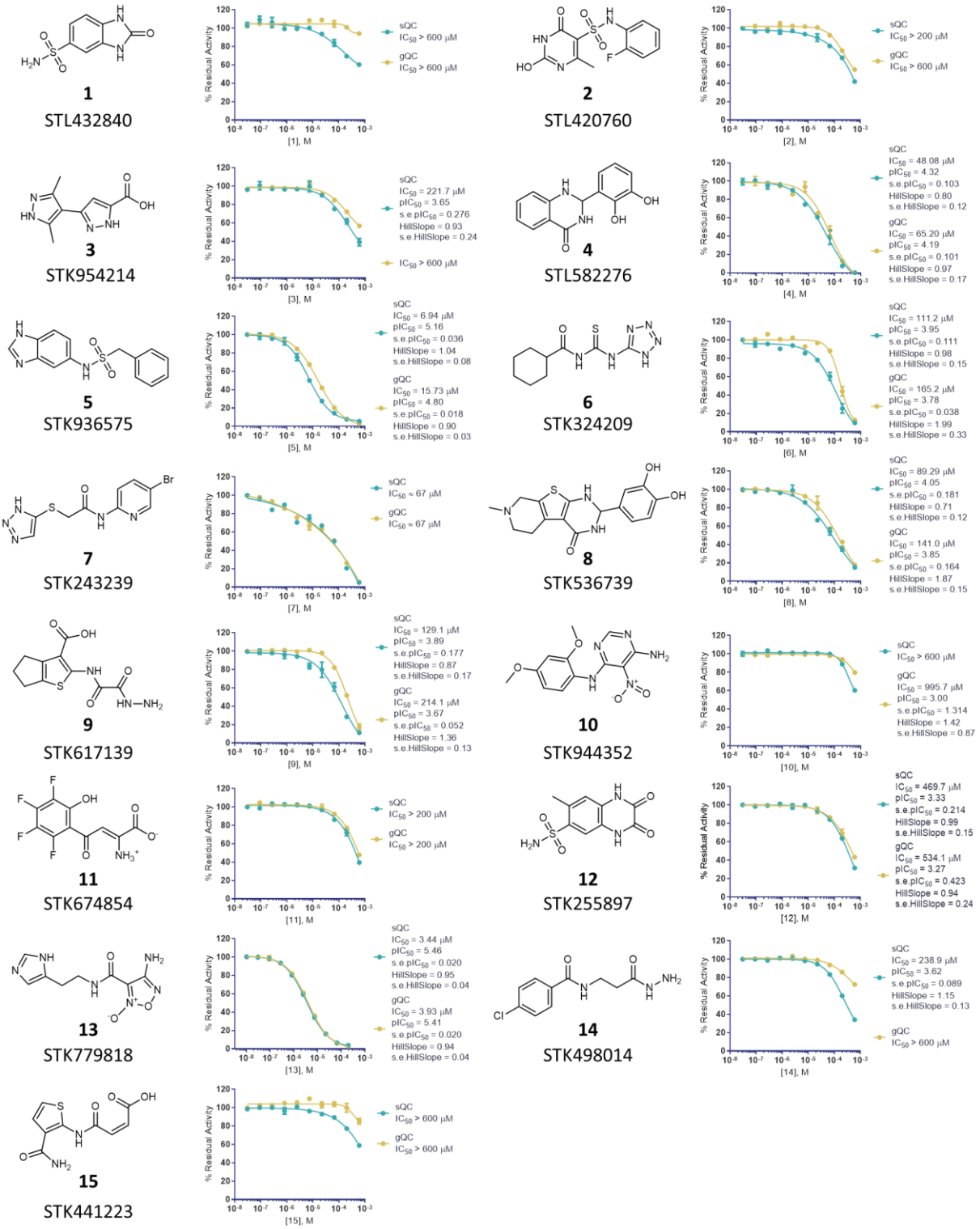
Supplementary Fig. 8 The performance comparison of DiffPhore with docking tools. **a,b** Cumulative distribution plots illustrating the proportion of observations below each RMSD value for different methods on (**a**) the PDBBind test set and (**b**) PoseBusters set. **c** The top-1 success rates for DiffPhore and docking methods evaluated on the full set (all proteins) or new protein subset (new proteins, not included in the training set) of PDBBind test set and PoseBusters set.



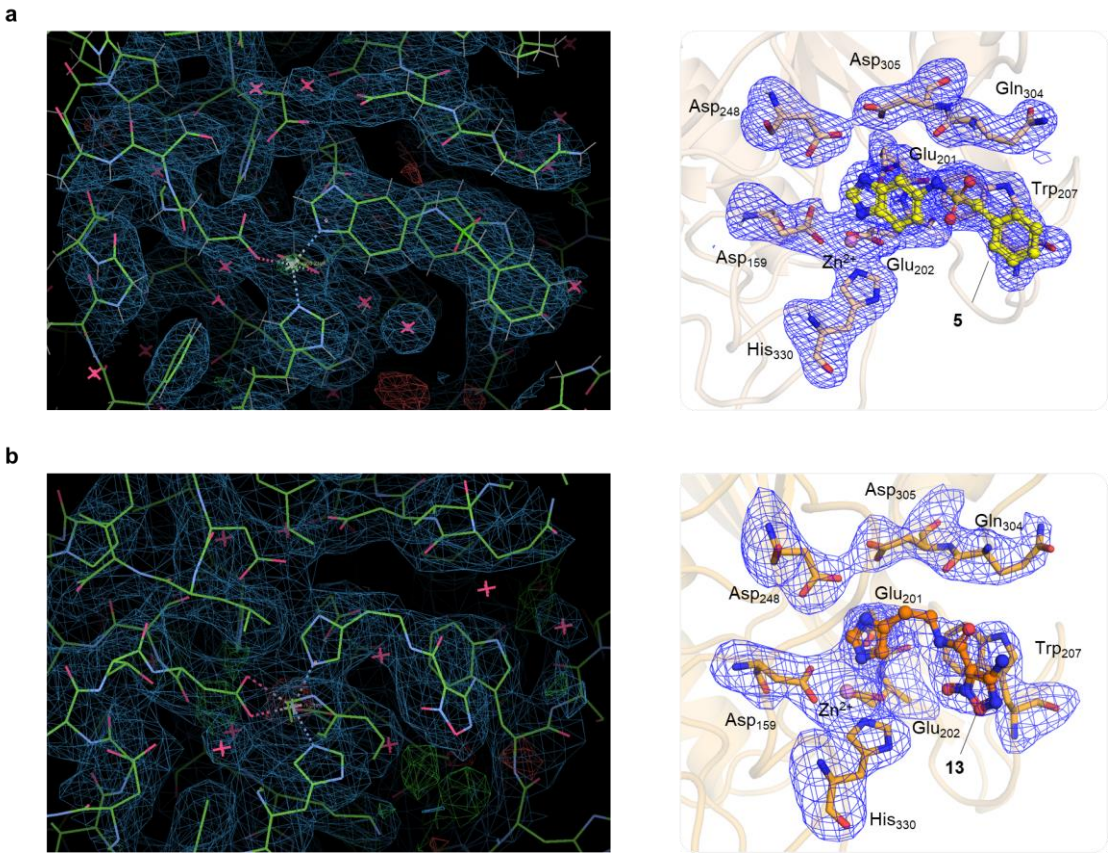
Supplementary Fig. 9 The impact of molecule flexibility and pharmacophore complexity on the prediction speed of DiffPhore and AncPhore. **a-c** Plots of running time versus (**a**) the number of heavy atoms, (**b**) rotatable bonds, and (**c**) pharmacophore features of the samples in PDBBind test set and PoseBuster set, revealing that unlike AncPhore, DiffPhore only experienced a modest reduction in speed when handling more flexible ligands. Data are presented as mean values +/- 95% confidence interval.



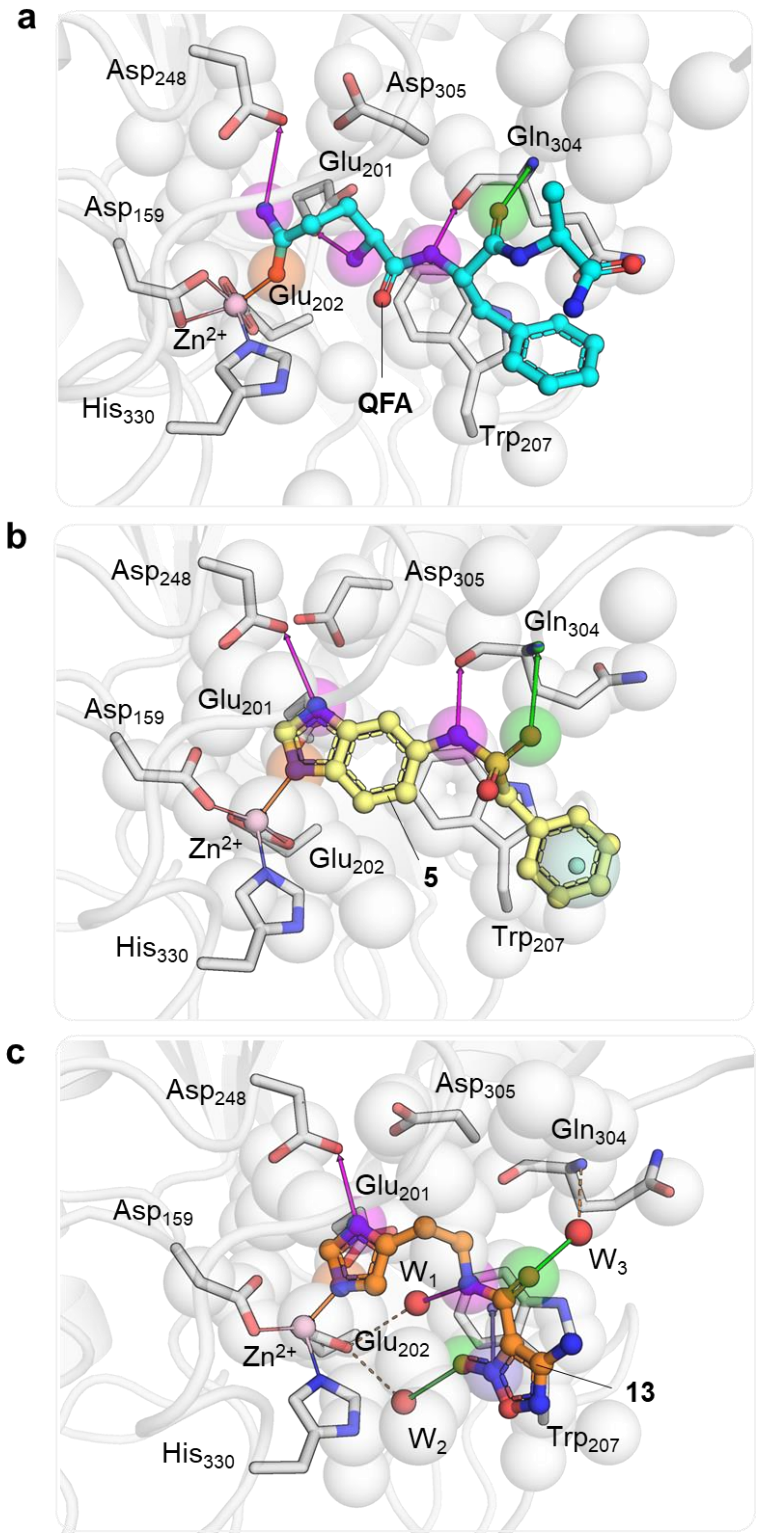
Supplementary Fig. 10 Comparison of enrichment factors for DiffPhore and the baseline methods in virtual screening for lead discovery. **a,b** Enrichment factors EF1% evaluated on (**a**) non-overlapping targets and (**b**) overlapping targets with the training set. EF1%: enrichment factor at 1%. Boxes are ranked based on their mean values, indicated by triangle markers. The '*' symbol denotes a statistically significant difference (unpaired one-sided student's t-tests, p-value < 0.05, n=14) between the baseline and DiffPhore ($DfScore1$). Exact p values are provided in the Source Data file. The boxes represent data distribution with center lines showing medians, box limits indicating the 25th and 75th percentiles, and whiskers extending to 1.5 times the interquartile range from the lower and upper quartiles.



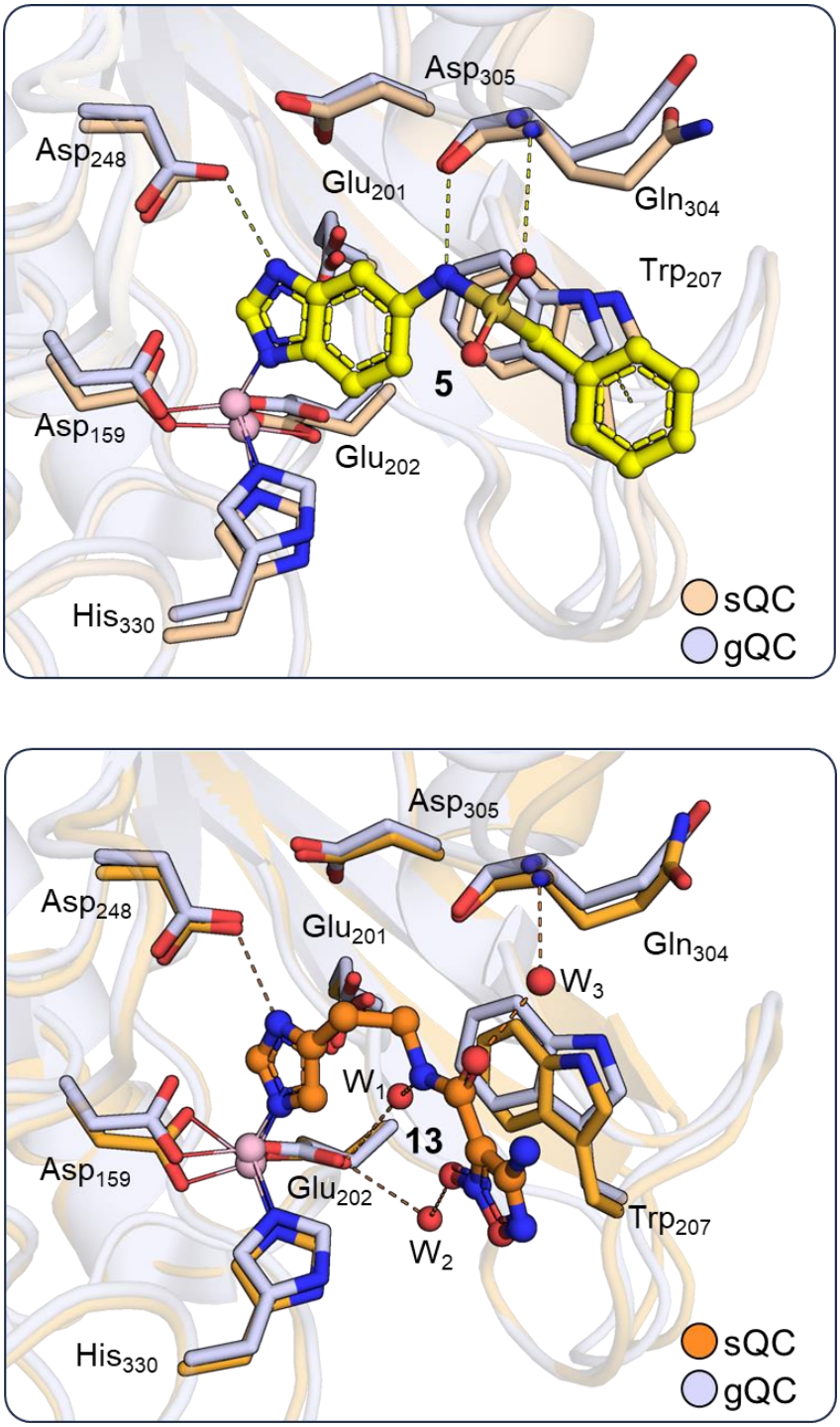
Supplementary Fig. 11 The IC_{50} curves of the 15 selected compounds against human sQC and gQC. All determinations are tested in triplicate. Data are presented as IC_{50}, pIC_{50} and s.e. pIC_{50}.



Supplementary Fig. 12 Electron density maps (2Fo-Fc, contour level: σ = 1, left) and omit maps (mF_o-DF_c, blue mesh, contour level: 3.0 σ, right) of the active sites for the (**a**) sQC:**5** (PDB code 9ISD) and (**b**) sQC:**13** (PDB code 9IVV) structures.



Supplementary Fig. 13 Comparison of the key pharmacophore features of QFA, **5**, and **13** binding with sQC. Views of the pharmacophore features derived from (**a**) sQC:**QFA**, (**b**) sQC:**5** (PDB code 9ISD), and (**c**) sQC:**13** (PDB code 9IVV) complexes, revealing their common pharmacophore features, *e.g.* those regarding zinc coordination and hydrogen bonding with the catalytic triad Asp_{248} and Gln_{304}.



Supplementary Fig. 14 Superimposition of (**a**) sQC:**5** (PDB code 9ISD) and (**b**) sQC:**13** (PDB code 9IVV) complex structures with the gQC structures (PDB code 3PB7)[10] indicates that compounds **5** and **13** are likely to bind to gQC through a mode of inhibition similar to that observed for sQC.

Supplementary Tables

Supplementary Table 1. The chemical properties of LigPhoreSet and CpxPhoreSet.

Chemical property	Mean (standard deviation)	
	LigPhoreSet	CpxPhoreSet
Number of heavy atoms	24.84 (5.03)	34.64 (22.20)
Number of rotatable bonds	4.40 (2.04)	12.79 (13.35)
Number of bonds	27.28 (5.63)	67.14 (46.34)
Number of rings	3.45 (0.91)	3.08 (1.70)
Number of hydrogen bond acceptors	5.97 (1.87)	10.27 (8.92)
Number of hydrogen bond donors	1.17 (1.06)	5.72 (6.72)
Molecular weight	351.68 (70.44)	496.65 (316.29)
cLogP	2.64 (1.27)	0.86 (4.07)
Topological polar surface area (TPSA)	74.01 (27.23)	158.18 (150.55)

Supplementary Table 2. Average counts and standard deviations for each pharmacophore feature in LigPhoreSet and CpxPhoreSet.

Pharmacophore feature count	Mean (standard deviation)	
	LigPhoreSet	CpxPhoreSet
Number of pharmacophore features	5.66 (1.28)	6.31 (2.72)
Number of exclusion spheres (EX)	95.98 (30.49)	80.08 (22.83)
Number of pharmacophore features with direction	4.42 (1.40)	4.60 (2.60)
Number of pharmacophore features without direction	1.23 (0.96)	1.73 (1.37)
Number of metal coordination (MB)	0.80 (0.55)	0.22 (0.67)
Number of hydrogen bond donor (HD)	0.46 (0.63)	1.58 (1.56)
Number of hydrogen bond acceptor (HA)	1.89 (1.13)	2.11 (1.66)
Number of positively charged center (PO)	0.01 (0.12)	0.16 (0.41)
Number of negatively charged center (NE)	0.02 (0.14)	0.21 (0.50)
Number of aromatic ring (AR)	0.61 (0.71)	0.32 (0.69)
Number of hydrophobic feature (HY)	1.00 (0.91)	1.31 (1.33)
Number of covalent bond (CV)	0.20 (0.40)	0.05 (0.21)
Number of cation-π interaction feature (CR)	0.58 (0.70)	0.36 (0.74)
Number of halogen bond (XB)	0.08 (0.30)	0.02 (0.14)

Supplementary Table 3. Calibrated conformation sampling algorithm.

Algorithm: Calibrated conformation sampling for training DiffPhore

Input: Ligand-pharmacophore pairs $\{(G_{l,o},\ G_p)\}$ from CpxPhoreSet

for epoch ← 1 **to** 800 **do**

  foreach $G_{\mathrm{l,o}},\ G_p$ **do**

    Sample $P \sim Uni([0,1])$

    $P_{\mathrm{epoch}} = EDSched(epoch)$

    Sample $t \sim Uni([0,\ 1])$

    if $P > P_{\mathrm{epoch}}$, **then**

      Sample $\Delta\mathbf{r}_{0\to t},\ \Delta\mathbf{R}_{0\to t},\ \Delta\boldsymbol{\theta}_{0\to t}\ from\ p_t^{tr}(\cdot|0),\ p_t^{rot}(\cdot|0),\ p_t^{tor}(\cdot|0)$

      Compute $\mathbf{x}_t \leftarrow A\big((\Delta\mathbf{r}_{0\to t},\ \Delta\mathbf{R}_{0\to t},\ \Delta\boldsymbol{\theta}_{0\to t}),\ \mathbf{x}_0\big)$

      Compute $\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t \leftarrow \nabla p_t^{tr}(\Delta\mathbf{r}_{0\to t}|0),\ \nabla p_t^{rot}(\Delta\mathbf{R}_{0\to t}|0),\ \nabla p_t^{tor}(\Delta\boldsymbol{\theta}_{0\to t}|0)$

      Predict scores $\widehat{\boldsymbol{\alpha}}_t, \widehat{\boldsymbol{\beta}}_t, \widehat{\boldsymbol{\gamma}}_t = DiffPhore\big(\hat{G}_{l,t},\ G_p,\ t\big)$

      Take optimization step on loss

        $\mathcal{L} = \|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t\|^2 + \left\|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t\right\|^2 + \|\widehat{\boldsymbol{\gamma}}_{\mathrm{t}} - \boldsymbol{\gamma}_t\|^2$

    else

      $t' = t + \Delta t \quad \left(\Delta\mathrm{t} = \frac{1}{\mathrm{N}}, here\ N\ refers\ to\ the\ denoising\ steps, default\ 20\right)$

      Sample $\Delta\mathbf{r}_{0\to t'},\ \Delta\mathbf{R}_{0\to t'},\ \Delta\boldsymbol{\theta}_{0\to t'}\ from\ p_{t'}^{tr}(\cdot|0),\ p_{t'}^{rot}(\cdot|0),\ p_{t'}^{tor}(\cdot|0)$

      Compute $\mathbf{x}_{t'} \leftarrow A\big((\Delta\mathbf{r}_{0\to t'},\ \Delta\mathbf{R}_{0\to t'},\ \Delta\boldsymbol{\theta}_{0\to t'}),\ \boldsymbol{x}_0\big)$

      Predict scores $\widehat{\boldsymbol{\alpha}}_{t'}, \widehat{\boldsymbol{\beta}}_{t'}, \widehat{\boldsymbol{\gamma}}_{t'} = DiffPhore\big(\hat{G}_{l,t'},\ G_p,\ t\big)$

      $\Delta\sigma_{tr}^2 = \sigma_{tr}^2(n/N) - \sigma_{tr}^2((n-1)/N), similarly\ for\ \Delta\sigma_{rot}^2, \Delta\sigma_{tor}^2$

      Sample z_{tr}, z_{rot}, z_{tor} from $\mathcal{N}(0, \Delta\sigma_{tr}^2),\ \mathcal{N}(0, \Delta\sigma_{rot}^2),\ \mathcal{N}(0, \Delta\sigma_{tor}^2)$

      $\Delta\mathbf{r}_{t'\to t} = \mathbf{r}_0 + \Delta\sigma_{tr}^2\widehat{\boldsymbol{\alpha}}_{t'} + z_{tr}$

      $\Delta\mathbf{R}_{t'\to t} = \boldsymbol{R}\big(\Delta\sigma_{rot}^2\widehat{\boldsymbol{\beta}}_{t'} + z_{rot}\big)$

      $\Delta\boldsymbol{\theta}_{t'\to t} = \Delta\sigma_{tor}^2\widehat{\boldsymbol{\gamma}}_{t'} + z_{tor}$

      Compute $\tilde{\mathbf{x}}_{\boldsymbol{t}} \leftarrow A\big((\Delta\mathbf{r}_{0\to t},\ \Delta\mathbf{R}_{0\to t},\ \Delta\boldsymbol{\theta}_{0\to t}),\ \mathbf{x}_{t'}\big)$

      $\Delta\boldsymbol{\theta}_{0\to t} = \Delta\boldsymbol{\theta}_{0\to t'} + \Delta\boldsymbol{\theta}_{t\to t'}$

      $\mathbf{x}_t^{tor} = A_{tor}(\mathbf{x}_0,\ \Delta\boldsymbol{\theta}_{0\to t})$

      $\mathbf{R},\ \mathbf{T} = Superimpose(\mathbf{x}_t^{tor},\ \tilde{\mathbf{x}}_{\boldsymbol{t}})$

      $\Delta\mathbf{r}_{0\to t} = \mathbf{T} + mean(\mathbf{x}_t^{tor})@R - mean(\mathbf{x}_t^{tor})$

      $\Delta\mathbf{R}_{0\to t} = \mathbf{R}$

      Compute $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\beta}}_{\boldsymbol{t}}, \tilde{\boldsymbol{\gamma}}_{\boldsymbol{t}} \leftarrow \nabla p_t^{tr}(\Delta\mathbf{r}_{0\to t}|0), \nabla p_t^{rot}(\Delta\mathbf{R}_{0\to t}|0), \nabla p_t^{tor}(\Delta\boldsymbol{\theta}_{0\to t}|0)$

      Predict Scores $\widehat{\boldsymbol{\alpha}}_t, \widehat{\boldsymbol{\beta}}_t, \widehat{\boldsymbol{\gamma}}_t = DiffPhore\big(\tilde{G}_{l,t},\ G_p,\ t\big)$

      Take optimization step on loss

        $\mathcal{L} = \|\widehat{\boldsymbol{\alpha}}_t - \tilde{\boldsymbol{\alpha}}_t\|^2 + \left\|\widehat{\boldsymbol{\beta}}_t - \tilde{\boldsymbol{\beta}}_{\boldsymbol{t}}\right\|^2 + \|\widehat{\boldsymbol{\gamma}}_t - \tilde{\boldsymbol{\gamma}}_{\boldsymbol{t}}\|^2$

    fi

Supplementary Table 4. Ablation study for DiffPhore model variants.[a]

Type Matching	Direction Matching	Success rate (%, RMSD<2 Å)	Ratio of $DfScore1$ > 0.7
Y	**Y**	**46.33**	**11.49**
Y	N	43.48	9.36
N	Y	-	-
N	N	43.84	10.43

[a] "Y" stands that the condition is considered, "N" means otherwise; "-" indicates unavailable data due to the model corresponding to the configuration failing to converge. All the models were trained on the CpxPhoreSet training set and evaluated on the CpxPhoreSet validation set. Number of initial poses = 1, and sampling steps = 20. The best results are highlighted in **bold**.

Supplementary Table 5. Ablation study for DiffPhore training schemes.[a]

Warm-up training	Refinement training	Calibrated conformation sampler	Full test set of PDBBind			Unseen protein subset of PDBBind		
			%<1	%<2	Med.	%<1	%<2	Med.
Y	**Y**	**Y**	**34.82**	**73.82**	**1.26**	**41.55**	**71.83**	**1.21**
N	Y	Y	33.98	68.80	1.36	36.62	67.61	1.37
Y	Y	N	34.82	71.03	1.33	37.32	64.79	1.34
N	Y	N	30.92	68.80	1.39	32.39	63.38	1.50
Y	N	N	18.36	43.79	2.16	22.86	50.0	2.02

[a] Given the wide diversity of ligands and pharmacophore features in LigPhoreSet, as well as their precise matching patterns, we utilized this set for initial training. CpxPhoreSet, based on real-world ligand-pharmacophore pairs, was employed for subsequent refinement training, with or without a calibrated conformation sampler. Warm-up training with LigPhoreSet Refinement training with CpxPhoreSet. “Y” stands that the condition is considered, “N” means otherwise. Number of initial poses = 40, sampling steps=20. Top-1 predictions ranked by $DfScore1$ are used for comparison.

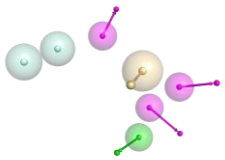
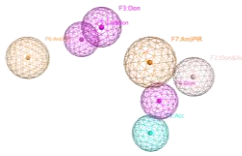
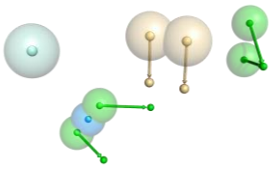
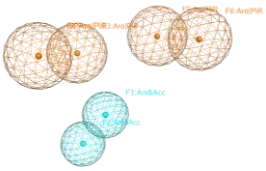
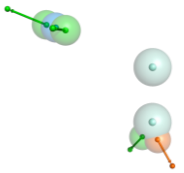
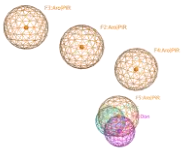
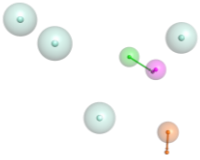
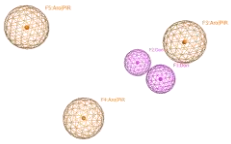
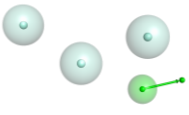
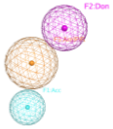
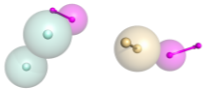
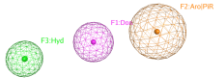
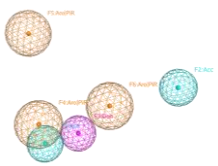
Supplementary Table 6. Evaluation of PoseBusters validity tests for each method on PDBBind test set and PoseBusters set.[a]

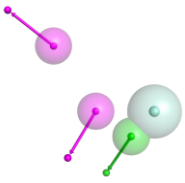
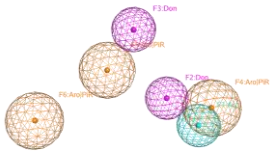
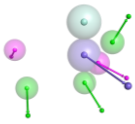
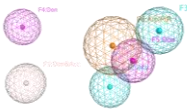
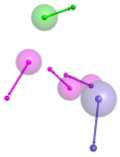
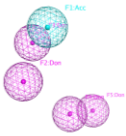
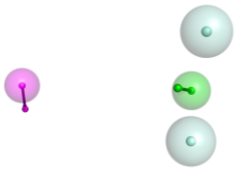
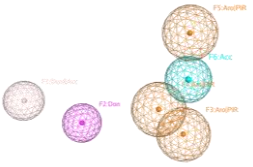
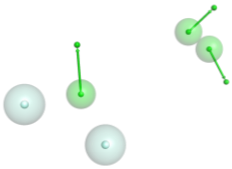
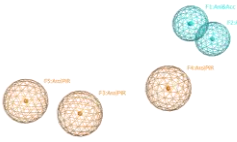
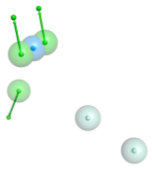
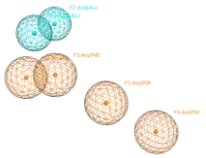
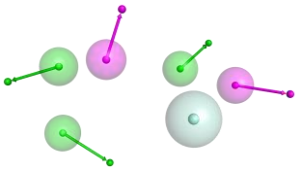
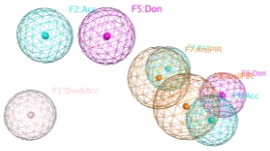
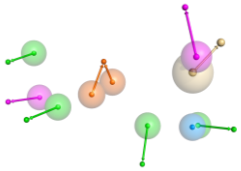
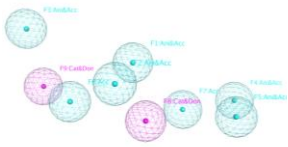
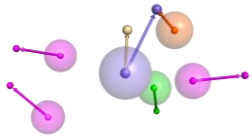
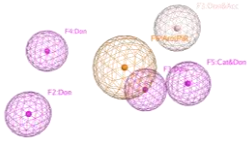
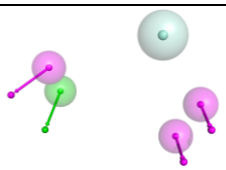
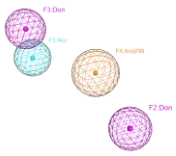
Dataset	Method	%RMSD<2Å & PB-Valid	%RMSD<2Å & PB-Valid (without protein)
PDBBind Test set	AncPhore(40,OB)	7.47	16.73
	AncPhore(10,OB)	6.61	14.79
	AncPhore(40,CF)	11.36	26.01
	AncPhore(10,CF)	12.45	26.91
	MOE(40, OB)	9.59	35.47
	MOE(10, OB)	8.14	31.98
	MOE(40, CF)	14.50	54.68
	MOE(10, CF)	12.99	44.11
	Uni-dock	**35.58**	36.81
	GNINA	26.84	42.17
	SMINA	21.08	30.42
	DiffPhore(40)	<u>31.80</u>	**72.48**
	DiffPhore(10)	21.41	<u>64.81</u>
PoseBusters	AncPhore(40, OB)	12.91	25.57
	AncPhore(10,OB)	15.18	29.27
	AncPhore(40,CF)	24.43	42.75
	AncPhore(10,CF)	25.75	43.84
	MOE(40, OB)	13.44	41.27
	MOE(10, OB)	11.32	34.67
	MOE(40, CF)	20.75	58.25
	MOE(10, CF)	16.51	50.71
	Uni-dock	34.47	34.47
	GNINA	**55.76**	61.93
	SMINA	45.67	48.95
	DiffPhore(40)	<u>53.68</u>	**86.70**
	DiffPhore(10)	42.04	<u>80.52</u>

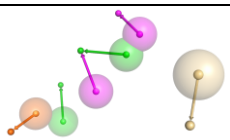
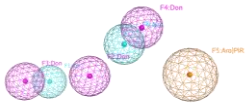
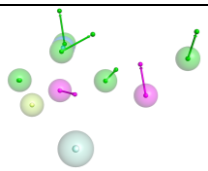
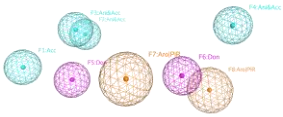
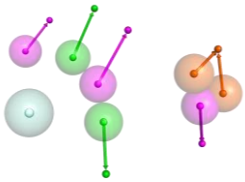
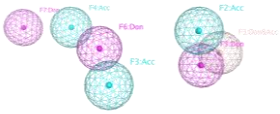
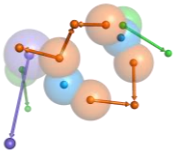
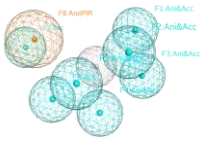
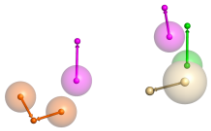
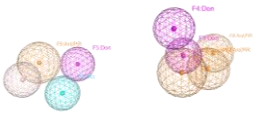
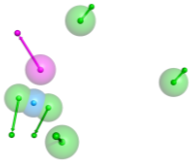
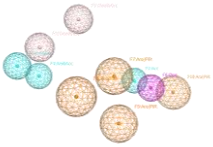
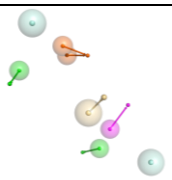
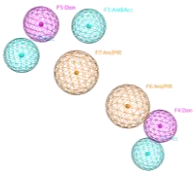
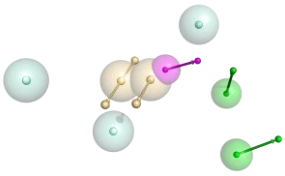
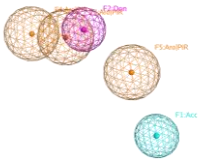
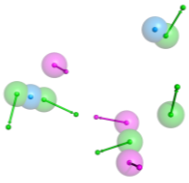
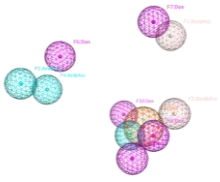
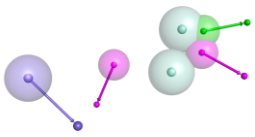
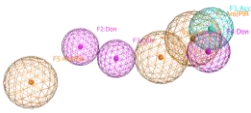
[a] The numbers in parentheses for DiffPhore, AncPhore and MOE represent the number of initial conformations. The abbreviations following the numbers in the parentheses denote the conformation tools for evaluation, where 'OB' refers to Openbabel and 'CF' refers to Conformator. The best and the second best results are highlighted in **bold** and <u>underlined</u>, respectively. The PoseBusters test suite is organized into three groups of tests: (1) Chemical validity and consistency; (2) Intramolecular validity; (3) Intermolecular validity. '%RMSD<2Å & PB-Valid' denotes the percentage of predicted poses that pass all PoseBusters validity tests and exhibit an RMSD of less than 2 Å compared to the ground truth poses. '%RMSD<2Å & PB-Valid (without protein)' denotes the percentage of

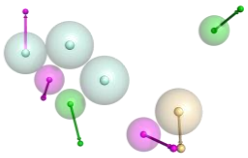
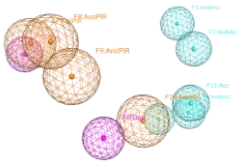
predicted poses that pass all PoseBuster validity tests, excluding the 'Intermolecular validity' test with protein, while also having an RMSD of less than 2 Å relative to the ground truth poses. The predicted poses of Autodock Vina are output in the PDBQT format files. These PDBQT files are converted to SDF file for PoseBusters validity tests, which throw a lot of errors about file parsing. The results for Autodock Vina might be problematic and thus not listed here.

Supplementary Table 7. The selected targets and corresponding pharmacophore models for evaluations of screening power for lead discovery.

Target	PDB code	Overlap with training set	Pharmacophore Model	
			DiffPhore/AncPhore[a]	MOE
ADRB2	6E67	No		
ALDR	2HV5	No		
CP2C9	5K7K	No		
CP3A4	3NXU	No		
DHI1	3FRJ	No		
DRD3	3PBL	No		
FAK1	3BZ3	No		

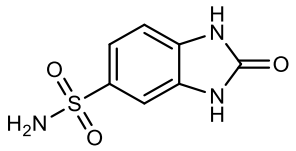
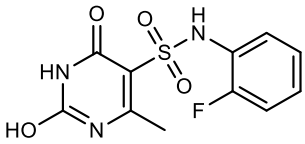
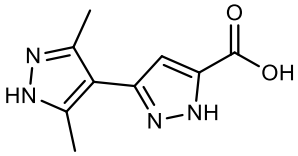
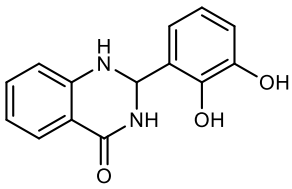
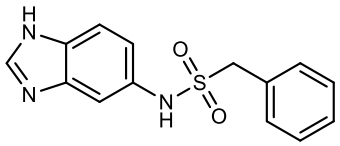
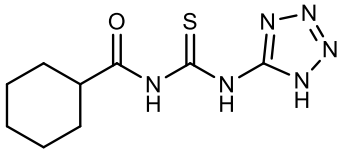
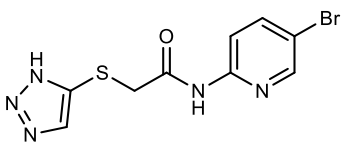
HXK4	3F9M	No		
KITH	2B8T	No		
NOS1	1QW6	No		
PGH1	2OYU	No		
PPARA	2P54	No		
PYRD	1D3G	No		
SAHH	1LI4	No		
ACE	6EN5	Yes		
ADA	1A4L	Yes		
BACE1	3L5D	Yes		

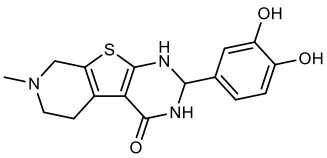
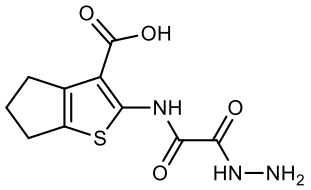
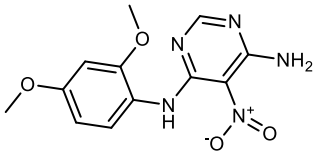
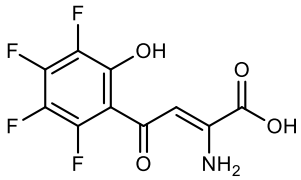
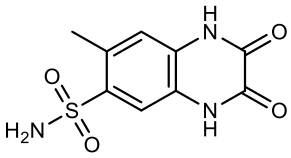
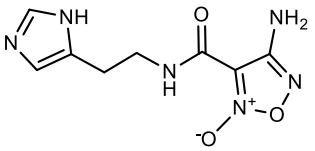
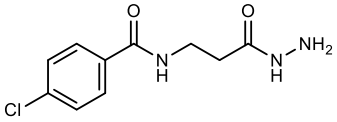
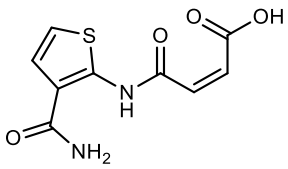
CAH2	5A6H	Yes		
CASP3	3KJF	Yes		
DEF	1IRU	Yes		
FPPS	1ZW5	Yes		
HDAC2	7KBG	Yes		
HMDH	1Q1L	Yes		
MMP13	1ZTQ	Yes		
PDE5A	2CHM	Yes		
PUR2	1NJS	Yes		
ROCK1	3V8S	Yes		

TYSY	1TSD	Yes		

[a] The color schemes for pharmacophore features in DiffPhore and AncPhore are detailed in Supplementary Fig. 1. For the pharmacophore definitions and color schemes used in MOE, please refer to the official MOE documentations. All the exclusion volumes are hidden for better visualization.

Supplementary Table 8. Chemical structures and inhibitory activities of compounds **1**-**15** (identified by DiffPhore-based virtual screening) against human sQC, gQC, and the auxiliary enzyme PGP-1.

Cpd. ID	Chemical Structure Vitas-M ID	IC_{50} (μM) / pIC_{50} / s.e. pIC_{50}		PGP-1 (Int%@100μM)
		sQC	gQC	
1	 STL432840	>600	>600	0 ± 1.53
2	 STL420760	>200	>600	14 ± 2.62
3	 STK954214	221.7 / 3.65 / 0.276	>600	9 ± 1.25
4[a]	 STL582276	48.08 / 4.32 / 0.103	65.2 / 4.19 / 0.101	63 ± 2.52
5	 STK936575	6.94 / 5.16 / 0.036	15.73 / 4.80 / 0.018	1 ± 1.23
6	 STK324209	111.2 / 3.95 / 0.111	165.2 / 3.78 / 0.038	60 ± 3.53
7	 STK243239	~ 67	~ 67	39 ± 1.03

8[a]	 STK536739	89.29 / 4.05 / 0.181	141 / 3.85 / 0.16	58 ± 2.31
9	 STK617139	129.1 / 3.89 / 0.177	214.1 / 3.67 / 0.052	0 ± 1.15
10	 STK944352	>600	995.7 / 3 / 1.314	6 ± 1.62
11	 STK674854	>200	>200	11 ± 1.73
12	 STK255897	469.7 / 3.33 / 0.214	534.1 / 3.27 / 0.423	18 ± 2.14
13	 STK779818	3.44 / 5.46 / 0.02	3.93 / 5.41 / 0.020	0 ± 0.62
14	 STK498014	238.9 / 3.62 / 0.089	>600	0 ± 1.85
15		>600	>600	0 ± 3.93

	STK441223

[a]Compounds **4** and **8** were tested as racemic mixtures. Note, since we only purchased 1~3 mg of these compounds, we did not conduct structural identification.

Supplementary Table 9. Crystallization conditions.

Crystal complex structure	Method	Protein sample composition	Crystallization reservoir condition	Experimental details
sQC:**5**	Co-crystallization	sQC in crystallization buffer[a], 2.14 mM compound **5**	0.2 M magnesium chloride, 15% (v/v) polyethylene glycol 4000, and 0.1 M Tris-HCl at pH 8.5	hanging drop vapor diffusion. 1:1 protein-to-reservoir ratio, 293K
sQC:**13**	Co-crystallization	sQC in crystallization buffer[a], 2.14 mM compound **13**	0.2 M magnesium chloride, 16% (v/v) polyethylene glycol 4000, and 0.1 M Tris-HCl at pH 8.5	hanging drop vapor diffusion. 1:1 protein-to-reservoir ratio, 293K

[a] The crystallization buffer contained 50 mM Tris-HCl, 150 mM NaCl, pH 8.0.

Supplementary Table 10. Data collection and refinement statistics for sQC:**5** (PDB code 9ISD) and sQC:**13** (PDB code 9IVV) complex structures.

Structure	sQC:**5**	sQC:**13**
PDB ID	9ISD	9IVV
Processing		
Radiation Source	SSRF Beamline BL18U1	
Space Group	*P* 1	*F* 41 3 2
Unit Cell	114.540	274.600
Dimensions	116.476	274.600
a, b, c (Å)	122.750	274.600
Unit Cell	96.21	90.00
Dimensions	114.92	90.00
α, β, γ (˚)	109.69	90.00
*Mol/ASU	12	1
Resolution Range (outer shell) (Å)	39.21-2.37 (2.44-2.37)	34.33-2.96 (3.04-2.96)
Number of Unique Reflections	208064	19022
Completeness (%)	97.8	99.9
$I/\sigma(I)$ (outer shell)	0.49	1.5
R_{merge} (outer shell)	1.8774	4.906
CC1/2	0.997	0.1964
Wilson B Factor (Å^2)	46.71	68.7
Refinement		
Overall B Factor (Å^2)	52.11	56.37
Protein B Factor (Å^2)	52.43	56.52
‡Ligand B Factor (Å^2) (occupancy)	42.26	55.62
Water B Factor (Å^2)	46.01	40.13
‡RMSD from Ideal Bond Length (Å)	0.024	0.023
RMSD from Ideal Angles (˚)	1.801	1.325
R_{work} (%)	18.91	20.60
R_{free} (%)	26.21	25.52

Supplementary Table 11. The detailed hyperparameters for DiffPhore.

Hyperparameters	Options
Learning rate	1-e5, 4e-5, 1e-4, 4e-4, **1e-3**
Batch size	10
Type match	**True**, False
Direction match	**True**, False
Using ligand hydrogens	True, **False**
Message passing layers (L)	3
Number of scalar features	20
Number of vector features	10
Using calibrated conformation sampler	**True**, False
Warm-up training	**True**, False
p_{max}	0.2, 0.4, **0.6**, 0.8
μ	**400**
c	**6**, 8, 10
Number of denoising steps	**20**

Note: The hyperparameters used in the final DiffPhore model are marked in **bold**. Other hyperparameters can be found in the source code repository.

Supplementary References

1. Dai, Q. et al. AncPhore: A versatile tool for anchor pharmacophore steered drug discovery with applications in discovery of new inhibitors targeting metallo-β-lactamases and indoleamine/tryptophan 2,3-dioxygenases. *Acta Pharm. Sin. B* **11**, 1931-1946 (2021).
2. Geiger, M. & Smidt, T. E. J. A. e3nn: Euclidean Neural Networks. Preprint at https://doi.org/10.48550/arXiv.2207.09453 (2022).
3. Ning, M. et al. Input Perturbation Reduces Exposure Bias in Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2301.11706 (2023).
4. Zhang, W. et al. Bridging the Gap between Training and Inference for Neural Machine Translation. Preprint at https://arxiv.org/abs/1906.02448 (2019).
5. Corso, G. et al. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. Preprint at https://doi.org/10.48550/arXiv.2210.01776 (2022).
6. Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **13**, 376-380 (1991).
7. Buttenschoen, M. et al. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chem. Sci.* **15**, 3130-3139 (2024).
8. Mou, J. et al. X-ray Structure-Guided Discovery of a Potent Benzimidazole Glutaminyl Cyclase Inhibitor That Shows Activity in a Parkinson's Disease Mouse Model. *J. Med. Chem.* **67**, 8730-8756 (2024).
9. Yung-Chi, C. & Prusoff, W. H. Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**, 3099-3108 (1973).
10. Huang, K.-F. et al. Structures of Human Golgi-resident Glutaminyl Cyclase and Its Complexes with Inhibitors Reveal a Large Loop Movement upon Inhibitor Binding, *J. Biol. Chem.* **286**, 12439-12449 (2011).