On the frustration to predict binding affinities from protein-ligand structures with deep neural networks.

Mikhail Volkov,[†] Joseph-André Turk,[≠] Nicolas Drizard,[≠] Nicolas Martin,[≠] Brice Hoffmann,[≠] Yann Gaston-Mathé[≠] and Didier Rognan[†]*

† Laboratoire d'innovation thérapeutique, UMR7200 CNRS-Université de Strasbourg, 74 route du Rhin, 67400 Illkirch, France.

≠ Iktos, 65 rue de Prony, 75017 Paris, France.

* To whom correspondence should be addressed (phone: +33 3 68 85 42 35, fax: +33 3 68 85 43 10, email: rognan@unistra.fr)

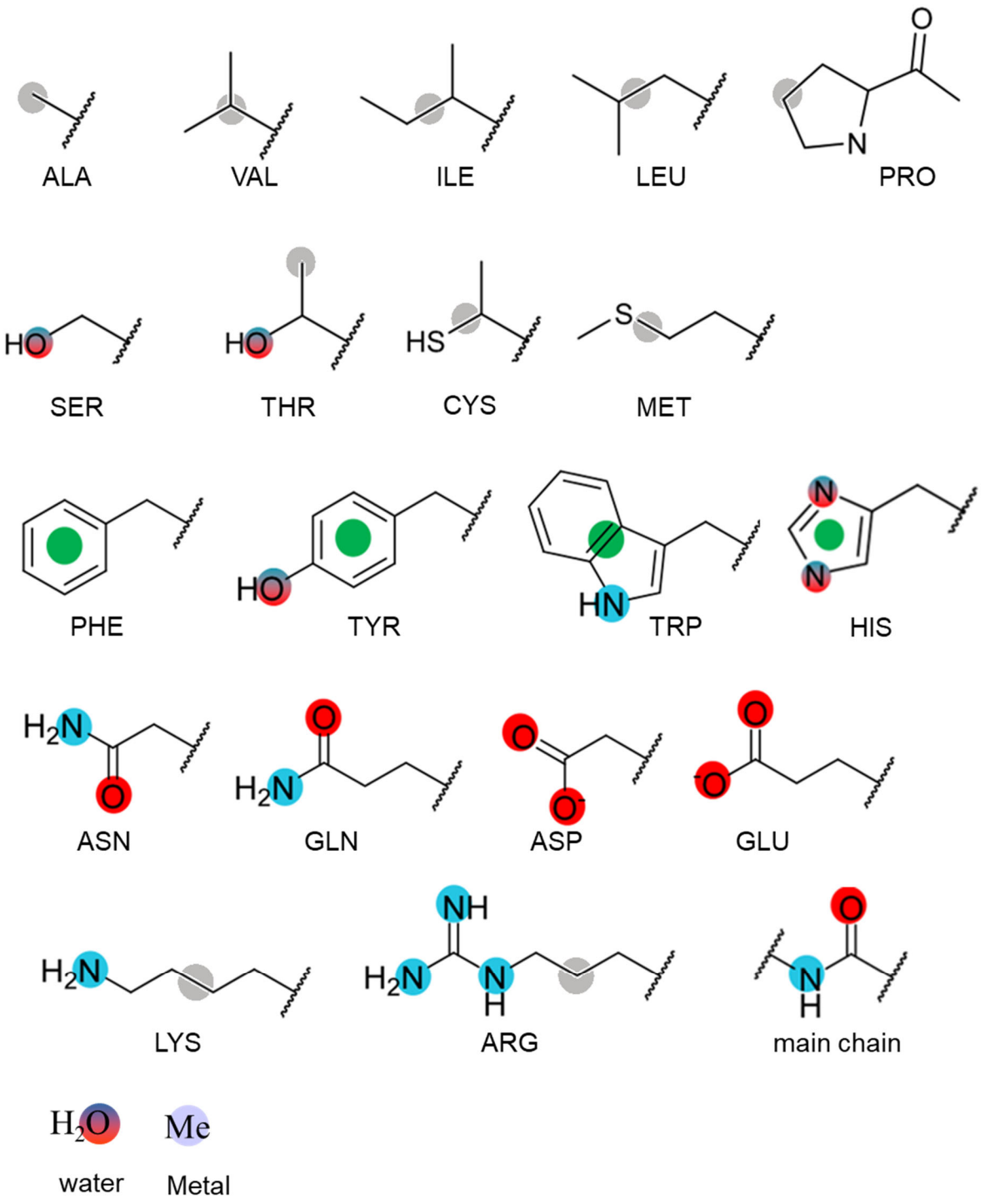


Figure S1. Location and pharmacophoric properties of protein pseudoatoms (grey, aliphatic; red, hydrogen-bond acceptor; cyan, hydrogen-bond donor; green, aromatic; metal-chelating, steel blue)

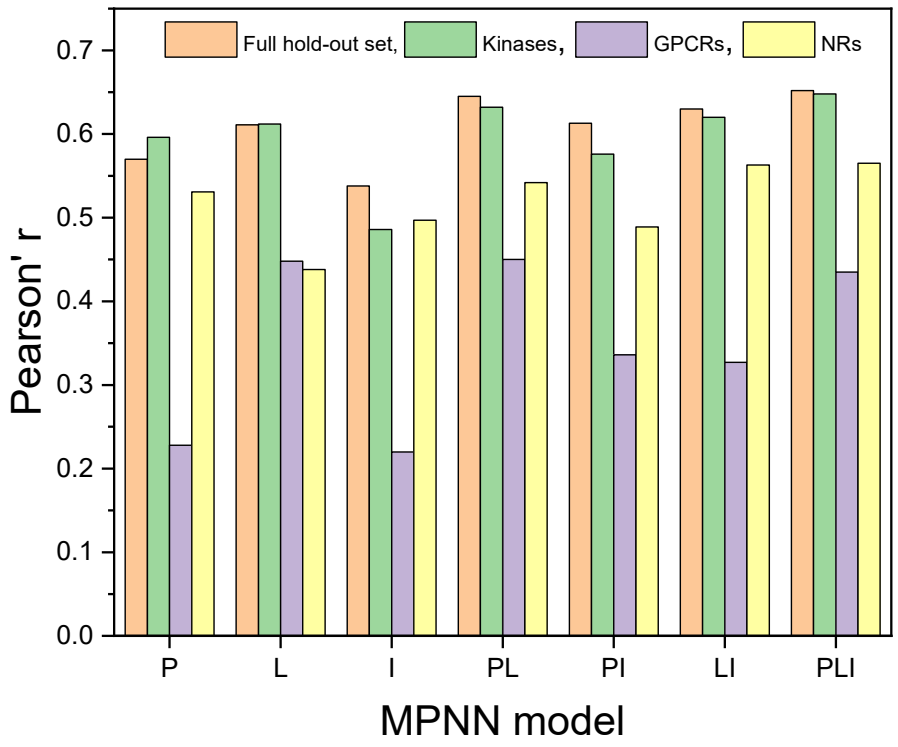


Figure S2. Performance of modular MPNN models in predicting affinities for specific target classes of the 2019 hold-out set. Mapping of protein target classes (GPCRs, G-protein-coupled receptor; NRs, Nuclear hormone receptors) from the Pharos database (https://pharos.nih.gov/) to PDB entries was performed using the Pharos-to-PDB code (https://github.com/ravila4/Pharos-to-PDB).

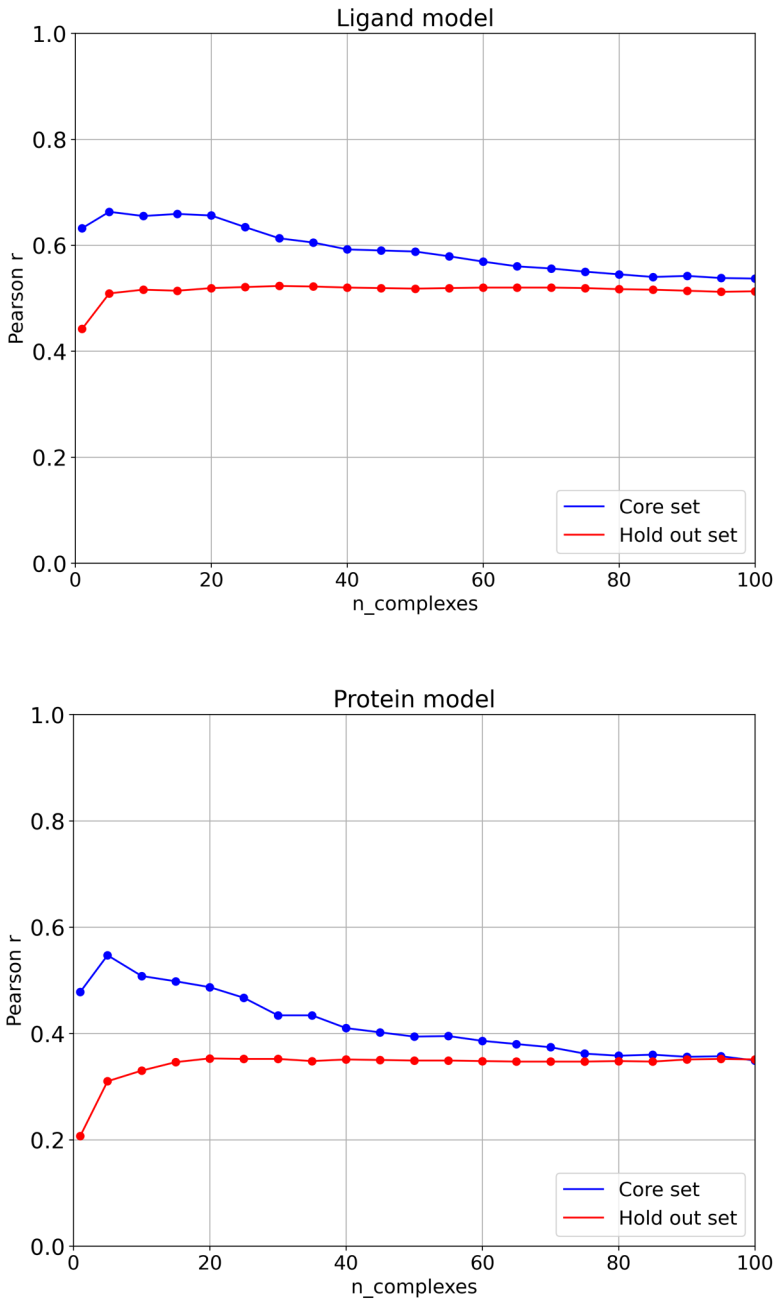
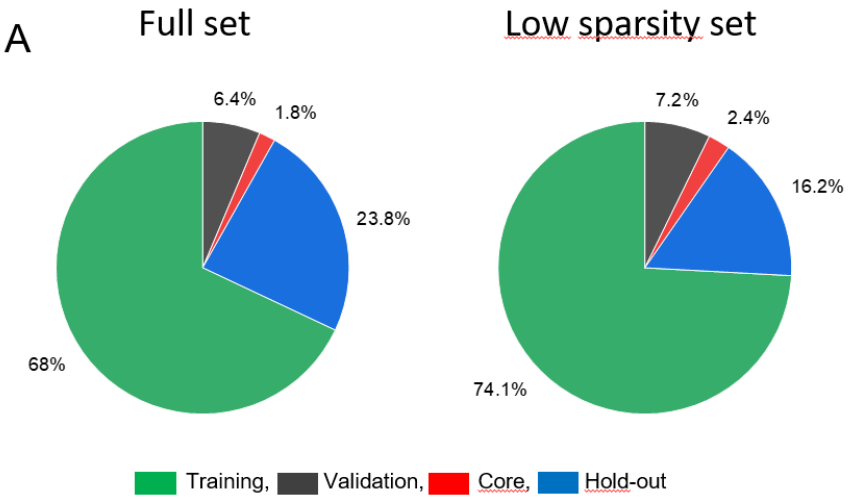


Figure S3. Influence of the number of closest ligands or proteins used to average binding affinities in the performance of simple memorization models, assessed by the Pearson r correlation coefficient.



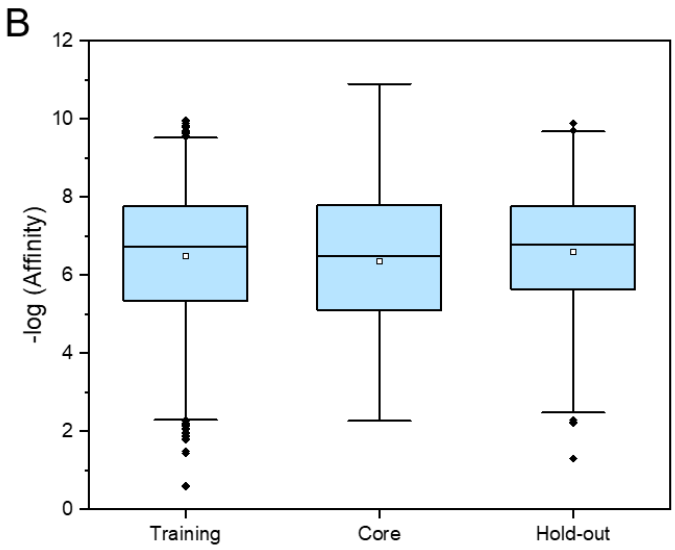


Figure S4. PDBbind low sparsity subset. **A)** Split into four subsets for training, validation and external test sets (core 2016, hold-out 2019), **B)** Distribution of experimental affinities for the training (n=1,505), core (n=49) and hold-out sets (n=329). The boxes delimit the 25th and 75th percentiles, and the whiskers delimit the 1st and 99th percentiles. The median and mean values are indicated by a horizontal line and a filled square in the box, respectively. Outliers are indicated by a diamond.

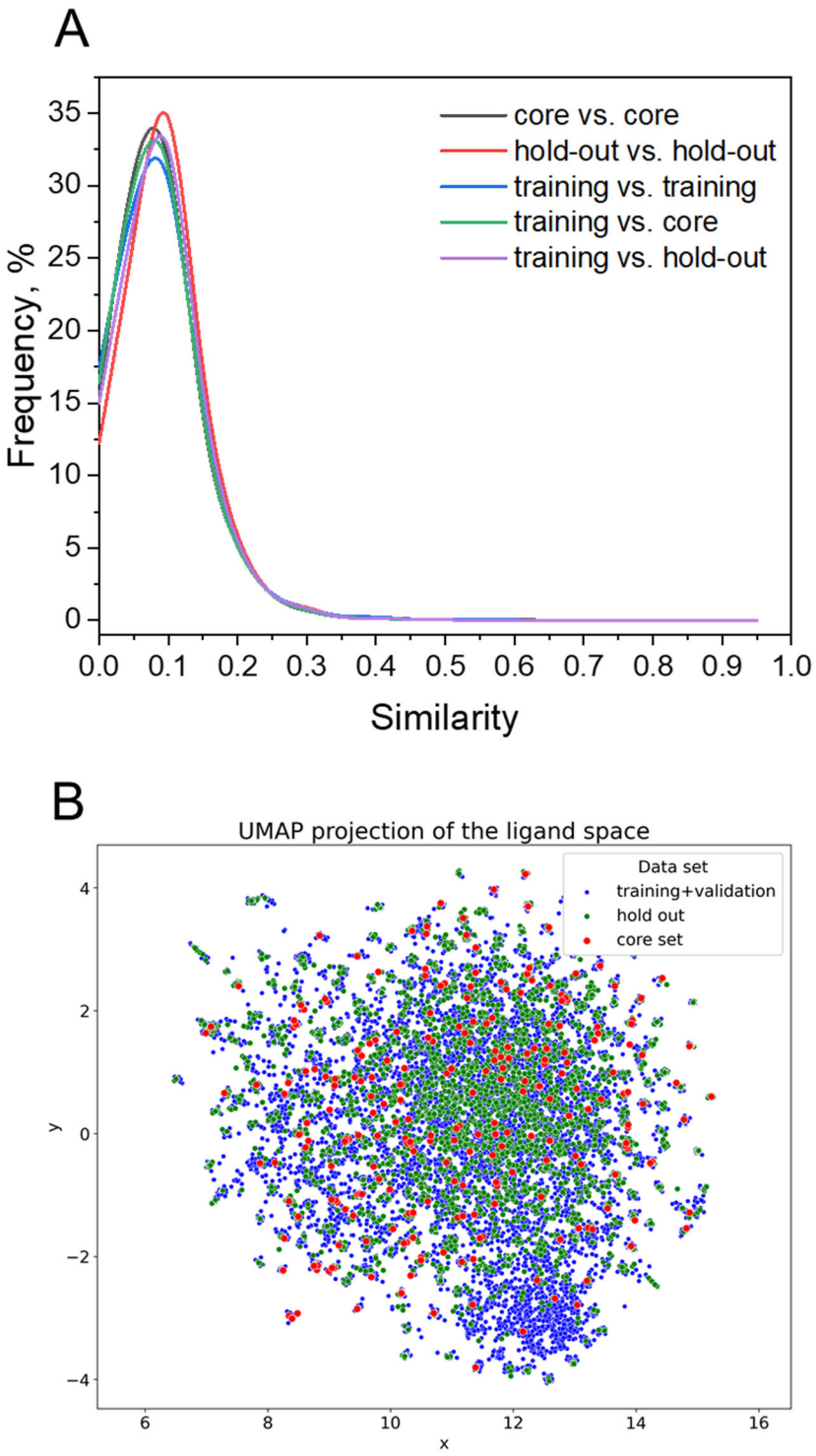


Figure S5. Chemical diversity of PDBbind ligands. **A)** Pairwise similarity of Murcko superstructures for ligands from the training, 2016 core and 2019 hold-out sets. The similarity is expressed by the Tanimoto coefficient computed from ECFP4 fingerprints. **B)** Uniform Manifold Approximation and Projection (UMAP)of PDBbind ligands, performed in umap-learn 0.5.3 with a number of neighbours of 30 and a dice distance metric. The Morgan fingerprints of radius 2 (nBits=1024) were computed in rdkit v.2020.09.1.

Table S1. Structure-based deep neural networks to predict protein-ligand binding affinities.

Model	Type	Objects	Descriptor	Training set	Test set	Split	R*p*	RMSE	Reference
TNet-BP	CNN	PL	Topological fingerprints	PDBbind 2016 refined (n=3767)	PDBbind 2016 core (n=290)	PDBBind original	0.826	1.37	11
ACNN	CNN	P, L, PL	Atom type-labelled distances (Nat*25 atom types*12 closest neighbors)	PDBbind 2015 refined (n=2965)	PDBbind 2015 refined (n=741)	Temporal	0.727	-	12
Brendan	CNN	PL	3D grid (21*21*21 Å) * 256 bit SPLIF vector	PDBbind 2016 general (10000)	PDBbind 2016 general (1500)	Random	0.704	-	13
PotentialNet	GNN	P, L, PL	Protein-ligand graph	PDBbind 2007 refined (n=1095)	PDBbind 2007 core (n=195)	PDBBind original	0.822	1.39	14
K_{DEEP}	CNN	L, PL	1-Å 3D grid (25*25*25 Å) * 16 features	PDBbind 2016 refined (n=3767)	PDBbind 2016 core (n=290)	PDBbind original	0.820	1.27	15
Pafnucy	CNN	L, PL	1-Å 3D grid (21*21*21 Å) * 19 features	PDBbind 2016 general (11906)	PDBbind 2016 core (n=290)	PDBBind original	0.780	1.42	16
DeepATom	CNN	PL	1-Å 3D grid (25*25*25 Å) * 24 features	PDBbind 2016 refined (n=3390)	PDBbind 2016 core (n=290)	PDBbind original	0.807	1.32	18
DeepBindRG	CNN	PL	ligand (84) * protein (41) atom pair distances < 4 Å	PDBbind 2018 general (n=13500)	PDBbind 2018 general (n=925)	Random	0.593	1.50	21
OnionNet	CNN	PL	Atom type-labelled distances (Nat*25 atom types*12 closest neighbors)	PDBbind 2016 general (n=11906)	PDBbind 2016 core (n=290)	PDBbind original	0.816	1.28	22
RosENet	CNN	PL	voxelized Rosetta interaction energies + pharmacophoric descriptors	PDBbind 2016/2018 refined (n=4463)	PDBbind 2016 core (n=290)	PDBBind original	0.820	1.24	23
graphDelta	GNN	L, PL	One-hot encoded ligand atoms + protein environmental descriptors (373)	PDBbind 2018 general (n=8766)	PDBbind 2016 core (n=285)	PDBbind original	0.870	1.05	24
AK-Score	CNN	PL	id Kdeep	PDBbind 2016 refined (n=3772)	PDBbind 2016 core (n=285)	PDBBind original	0.827	1.22	25
SE-OnionNet	CNN	PL	1-Å grid (21*21*21)* 64 protein-ligand element distance counts	PDBbind 2018 general (n=11663)	PDBbind 2018 refined (n=463)	Random	0.853	1.59	27
Progessive multitask network	CNN	P, L, PL	ligand ECFP + Protein ECFP + Protein-Ligand SPLIF	PDBbind 2016 refined (n=3568)	PDBbind 2016 core (n=290)	PDBbind original	0.740	0.98	28
ACNN	CNN	P, L, PL	Atom type-labelled distances (Nat*25 atom types*12 closest neighbors)	PDBbind 2015 refined (n=3706)	PDBbind 2015 core (n=195)	PDBBind original	0.730	-	29
Pair	CNN	PL	protein-ligand distance pairs	PDBBbind 2018 refined (n=2675)	PDBBbind 2018 refined (n=891)	Random split	0.660	1.61	30
DEELIG	CNN	L, PL	Atomic model: 3D grid (10*10*10 Å) * 19 bits (atomic model); Composite model: 3D grid (10*10*10 Å) * 44 bits (pocket) + 14716 bits (ligand)	in-house set (n=4041)	PDBbind 2016 core (n=290)	Random 80/10/10	0.889	-	31
Interaction GraphNet	GNN	P, L, PL	independent GNN for intra and inter-molecular interactions	PDBbind 2016 general (n=10366)	PDBbind 2016 core (n=290)	PDBBind original	0.837	1.22	32

midlevel fusion	CNN+GNN	PL	CNN: 1-Å grid (48*48*48)* 19 atomic features; GNN: covalent (d < 1.5 Å) and non-covalent edges (1.5 < d < 4.5 Å)	Pdbbind 2016 general+refined (13283)	PDB2016 core set (n=290)	PDBBind original	0.810	1.31	33
SMPLIP	RF+ CNN	L, PL	IFP (140) + interaction distances (140) + SMF descriptors (2282)	Pdbbind 2016 general+refined (13283)	PDB2016 core set (n=290)	PDBBind original	0.770	1.51	34
OctSurf	CNN	PL	1-Å 3D grid (64*64*64 Å) * 24 features/octant	PDBbind 2018 general (n=16126)	PDBbind 2016 core (n=285)	PDBBind original	0.793	1.45	35
BAPA	CNN	PL	Protein-ligand interaction descriptors + 6 Vina terms	PDBbind 2016 refined (n=3689)	PDBbind 2016 core (n=285)	PDBBind original	0.819	1.31	36
APMNet	GNN+GNN	P, L	75 DeepChem atomic features	PDBbind 2016 general (n=11844)	PDBbind 2016 core (n=290)	PDBBind original	0.815	1.27	37
GraphBAR	GNN	PL	13 features * 200 protein-ligand atoms	PDBbind 2016 general (n=11146)	PDBbind 2016 core (n=290)	PDBBind original	0.764	1.44	38

Table S2. Geometric rules to define protein-ligand non-covalent interactions.

Interaction	Rule 1[a]	Rule 2[b]
H-bond	$\left\|\overrightarrow{DA}\right\| \le 3.5\mathring{A}$	$\left\langle \overrightarrow{DH}, \overrightarrow{HA} \right\rangle \in \left[\frac{-\pi}{4}, \frac{\pi}{4} \right]$
Ionic	$\left\|\overrightarrow{+-}\right\| \le 4.0\mathring{A}$	
Hydrophobe	$\left\|\overrightarrow{Y_1Y_2}\right\| \le 4.5\mathring{A}$	
Aromatic (Face to face)	$\left\|\overrightarrow{ac_1ac_2}\right\| \le 4.0\mathring{A}$	$\left\langle \overrightarrow{n_1}, \overrightarrow{n_2} \right\rangle \in \left[\frac{-\pi}{6}, \frac{\pi}{6} \right]$
Aromatic (Edge to face)	$\left\|\overrightarrow{ac_1ac_2}\right\| \le 4.0\mathring{A}$	$\left\langle \overrightarrow{n_1}, \overrightarrow{n_2} \right\rangle \in \left[\frac{\pi}{6}, \frac{5\pi}{6} \right]$
pi-cation	$\left\|\overrightarrow{ac+}\right\| \le 4.0\mathring{A}$	$\left\langle \overrightarrow{n}, \overrightarrow{ac+} \right\rangle \in \left[\frac{-\pi}{6}, \frac{\pi}{6} \right]$
Metal	$\left\|\overrightarrow{MA}\right\| \le 2.8\mathring{A}$	

[a] D: H-bond donor; A: H-bond acceptor; +: cation; -:anion; Y: hydrophobe; ac: geometric center of an aromatic ring; M: metal.

[b] H: hydrogen; n: normal to the aromatic ring.