

Supplementary Material for

Large-scale binding ligand prediction by improved patch-based method Patch-Surfer2.0

Xiaolei Zhu¹, Yi Xiong¹, and Daisuke Kihara^{1,2,*}

¹Department of Biology, Purdue University, West Lafayette, IN 47906, USA.

²Department of Computer Science, Purdue University, West Lafayette, IN 47906, USA.

Contact: dkihara@purdue.edu

Table S1. The 117 types of ligands that have more than 5 entries in the binding pocket database.

Ligand ID	Number of Pockets	Number of Atoms (atmn)	Number of Rotatable single bonds(rotb)	logP	rotb/atmn
AMP	46	23	4	-1.521	0.174
ATP	44	31	8	-3.535	0.258
FAD	82	53	13	-2.692	0.245
FMN	49	31	7	-1.426	0.226
GLC	27	12	1	-2.643	0.083
HEM	146	43	8	6.102	0.186
NAD	39	44	11	-6.064	0.250
HEZ	16	8	5	0.6	0.625
BCN	6	11	6	-2.589	0.545
CIT	120	13	5	-1.983	0.385
PMP	7	16	4	-1.367	0.250
GSH	7	20	9	-4.971	0.450
ANP	33	31	8	-3.734	0.258
ARG	10	12	6	-5.294	0.500
C8E	8	21	18	2.417	0.857
CDP	6	25	6	-3.605	0.240
MBO	12	10	1	1.432	0.100
PLM	11	18	14	7.059	0.778
PLP	30	16	4	-0.762	0.250
12P	9	37	34	-3.196	0.919
BGC	30	12	1	-2.643	0.0833
MCT	7	9	0	1.418	0.000
ASP	10	9	3	-3.518	0.333
DUP	6	28	8	-4.699	0.286
FUC	7	11	0	-1.635	0.000
HIS	7	11	3	-5.039	0.273
NHE	18	13	4	-0.914	0.308
TYD	6	25	6	-3.107	0.240
DTT	27	8	3	-0.335	0.375

DTU	8	8	3	-0.335	0.375
017	6	38	12	4.323	0.316
ACO	11	51	20	-4.096	0.392
ACP	10	31	8	-3.481	0.258
ACR	8	44	9	-5.514	0.205
GDP	21	28	6	-3.691	0.214
MET	11	9	4	-2.239	0.444
MYR	6	16	12	6.048	0.750
A2G	12	15	2	-3.083	0.133
HEC	13	43	6	5.542	0.140
NDG	18	15	2	-3.083	0.133
TMP	9	21	4	-2.1	0.190
2PE	19	28	25	-2.584	0.893
GTP	10	32	8	-4.529	0.250
P6G	20	19	16	-1.972	0.842
GTT	9	20	9	-4.971	0.450
TLA	29	10	3	-2.486	0.300
NGA	11	15	2	-3.083	0.133
GAL	16	12	1	-2.643	0.0833
SF4	45	8	0	5	0.000
U5P	12	21	4	-2.757	0.190
BLA	6	43	11	4.818	0.256
GNP	10	32	8	-4.64	0.250
PG4	113	13	10	-1.564	0.769
PG5	11	12	9	-0.315	0.750
PGE	69	10	7	-1.36	0.700
UDP	15	25	6	-3.764	0.240
TSU	10	11	1	-0.622	0.0909
SIA	10	21	5	-4.131	0.238
SIN	14	8	3	-0.655	0.375
TAM	8	11	6	-1.304	0.545
XYP	7	10	0	-2.216	0.000
COA	35	48	18	-4.444	0.375
FLC	33	13	5	-4.335	0.385
LEU	9	9	3	-1.382	0.333
TRP	9	15	3	-1.08	0.200
TRS	112	8	3	-4.618	0.375
BOG	17	20	9	1.432	0.450
EPE	71	15	5	-3.315	0.333
NAG	308	15	2	-3.083	0.133
NAP	37	48	13	-6.145	0.271
P33	15	22	19	-2.176	0.864
PE4	8	24	21	-1.277	0.875
PEB	6	43	12	5.1	0.279
PEG	158	7	4	-1.156	0.571

MES	110	12	3	-4.076	0.250
ADN	15	19	2	-0.854	0.105
MLA	34	7	2	-0.925	0.286
MLI	33	7	2	-3.49	0.286
MLT	18	9	3	-1.57	0.333
PG6	6	18	15	-0.723	0.833
TPP	8	26	8	-4.753	0.308
NCO	10	7	0	-5	0.000
AKG	8	10	4	-1.485	0.400
GLO	6	12	5	-2.643	0.417
GLU	12	10	4	-3.248	0.400
UD1	6	39	10	-5.179	0.256
15P	6	104	101	-5.426	0.971
1PE	43	16	13	-1.768	0.813
1PG	6	17	14	-1.245	0.824
1PS	7	13	4	-5.644	0.308
ADE	9	10	0	0.235	0.000
ADP	65	27	6	-2.528	0.222
CMP	12	22	1	-1.709	0.0455
MPD	182	8	2	0.492	0.250
MPO	7	13	4	-2.458	0.308
SUC	17	23	5	-3.745	0.217
CXS	10	14	5	-0.643	0.357
F3S	7	7	0	5	0.000
IHP	6	36	12	-5.547	0.333
POP	11	9	2	-4.922	0.222
APC	9	31	8	-3.481	0.258
APR	8	36	9	-3.409	0.250
B30	6	25	5	-1.87	0.200
B3P	11	19	12	-3.524	0.632
BTB	25	14	8	-2.76	0.571
CHT	7	7	2	-4.236	0.286
MRD	56	8	2	0.492	0.250
PRP	6	22	7	-3.817	0.318
SAH	27	26	7	-2.773	0.269
SAM	23	27	7	-4.143	0.259
BEN	16	9	1	0.323	0.111
BEZ	24	9	1	1.848	0.111
MAL	7	23	4	-4.45	0.174
MAN	33	12	1	-2.643	0.0833
NDP	24	48	13	-4.229	0.271
PQQ	6	24	3	0.235	0.125
SNG	6	16	3	-1.808	0.188

Procedure applied for selecting the non-redundant ligand binding pocket database

A non-redundant database of pockets with bound ligands was constructed based on the Protein-Small-Molecule Database http://compbio.cs.toronto.edu/psmdb/downloads/CPLX_25_0.85_7HA.list (PSMDB) (Wallach & Lilien, 2009). First, 5,438 protein-ligand complexes selected from PDB were obtained from PSMDB. Multiple ligands in the same pocket were united if they are closer than 1.4 Å, which indicates that heavy atoms are forming a covalent bond. The cutoff value of 1.4 Å was determined by considering the lengths of covalent bonds between carbon (C), nitrogen (N), and oxygen (O). For example, the length of a single bond of C-C is on average 1.54 Å, while the length of double and triple bonds between two carbons are 1.34 Å and 1.20 Å, respectively, and the average bond length for a carbon to carbon, nitrogen, or oxygen including single, double, and triple bonds are 1.32 Å.

Small united ligands with less than seven atoms were discarded, so that ions and small ligands, such as SO_4^{2-} are discarded. The seven atoms is the cutoff used by the PSMDB. Ligand-protein pairs where a ligand is covalently bound to a protein were also removed using a distance cutoff of 1.4 Å. In cases that multiple (united) ligands exist in a protein pocket, they are treated as a group if they are closer than 4.5 Å. 4.5 Å is a standard cutoff value used to define heavy atom contacts used in computational structural studies of proteins. It is larger than 3.5 Å used below but we observed that two ligands are consistently co-localize once they are observed closer than 4.5 Å in one of pockets. Grouped ligands are removed if they are not binding to a protein, i.e. if none of their heavy atoms is closer than 3.5 Å to any heavy atom of the protein. 3.5 Å is roughly the distance between two van der Waals radius of heavy atoms. This procedure yielded 9,361 pockets. Subsequently, we further removed redundant pockets that come from pockets in homo-multimers using two criteria: Pockets from homo-multimers in the same PDB file are considered as redundant and removed, keeping only one of them if the proteins have globally similar structures (an RMSD of less than 3.0 Å) and also the pockets share more than 80% of their residues. 3.0 Å RMSD and 80% identity are cutoff values we determined by examining many ligand binding pockets. If these two conditions are met, none pockets were observed to be substantially dissimilar in the shape and interactions with ligand molecules. Ligand binding residues were identified using a distance cutoff of 5.0 Å. The whole procedure resulted in 6,547 pockets. 5.0 Å is a commonly used cutoff value to determine protein-protein and protein-ligand interactions. Each ligand-pocket entry is classified into a ligand type, e.g. ATP, FAD, etc. to be able to perform binding ligand prediction. In case multiple ligands exist in a pocket, a larger ligand is selected as the representative if the ratio of the number of heavy atoms of the ligand is over 0.5 (i.e. half) to that of all the ligands. Finally, we identified 2,444 different main ligand types in the 6,547 pockets. 117 ligand types have more than five binding pockets in the dataset.

Reference:

The protein-small-molecule database, a non-redundant structural resource for the analysis of protein-ligand binding. Izhar Wallach and Ryan Lilien, *Bioinformatics*, 25: 615-620 (2009)

Table S2. Accuracy after excluding flexible ligands with different flexibility ratio.

Table 3 in the main manuscript shows the results after removing flexible ligands that have a flexibility ratio over 0.8. Here we show results after further removing flexible ligands with a flexibility ratio over 0.7, 0.6, and 0.5.

Removed ligands		Top 5	Top 10	Top 15	Top 20	Top25
Flexibility ratio cutoff	# of ligands removed					
-	0	0.234	0.417	0.526	0.592	0.642
0.8	10	0.252	0.452	0.567	0.635	0.683
0.7	14	0.250	0.454	0.571	0.639	0.686
0.6	17	0.257	0.461	0.577	0.643	0.689
0.5	21	0.264	0.472	0.588	0.652	0.695

The accuracy listed as the top row, the results without removing any ligand, is the same as the one listed at the top of Table 2 (the results of the 110 ligands). The second row is the same as the top row of Table 3.

Table S3. A set of holo and apo proteins used in Figure S3 and Table 4.


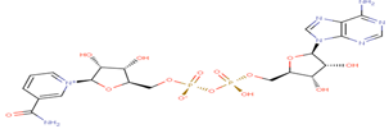
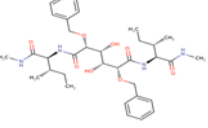
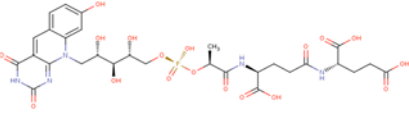
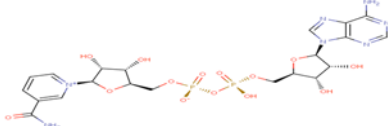
Holo protein	Binding ligand	Target ligand rank by holo protein	Apo protein	Target ligand rank by apo protein
1xqh_AB	SAH	17	1h3i_AB	20
2cch_A	ATP	3	1pw2_A	5
2fh6_A	GLC	10	2fgz_A	2
2fh8_A	BGC	3	2fgz_A	1
2gsd_A	NAD	3	2nac_A	2
2hq2_A	HEM	13	1u9t_A	3
2jhf_A	NAD	4	8adh_A	4
2phn_AB	GDP	9	2g9i_AB	9
2vjq_AB	EPE	26	1p5h_AB	37
3a6t_A	SUC	36	3a6s_A	21
3aar_A	ANP	14	3aap_A	13
3f9m_A	GLC	10	1v4t_A	2
3fuu_A	ADN	-	3fuv_A	18
3fxx_A	ANP	2	2gsf_A	4
3gru_A	AMP	14	3fuv_A	7
3hbn_A	UDP	14	3hbm_A	27
3lss_A	ATP	4	3lsq_A	2
3m0f_AB	GSH	-	3lxt_AB	10
3n5o_AB	GTT	-	3lg6_AB	-
1g5t_A	ATP	5	1g5r_A	3
1h6m_A	NAG	21	1sf4_A	23
1x2t_B	1PE	-	1x2w_B	28
2vlh_A	PGE	21	2ez2_A	15

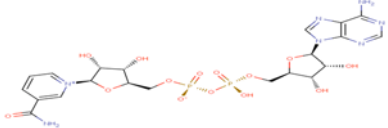
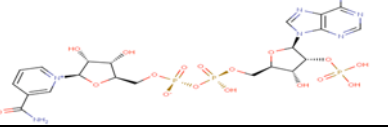
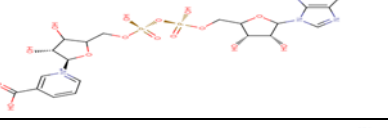
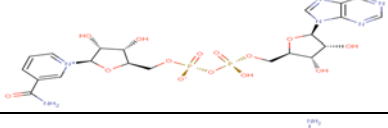
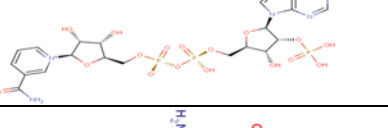
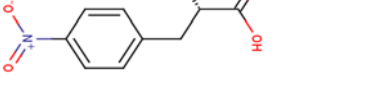
2vlh_A	P33	18	2ez2_A	6
3bpw_A	PEG	5	2q8l_A	7
3fek_A	PEG	28	2fs6_A	10
3gw8_A	PG4	25	3ezn_A	5
3gw8_A	PG4	6	3ezn_A	6
3gw8_AB	PG4	41	3ezn_AB	11
3lss_A	MLI	32	3lsq_A	-
3lss_A	MLI	20	3lsq_A	6
3mbm_AB	TRS	20	3f0e_AB	22

These apo proteins are identified as follows: For the proteins in the non-redundant binding pocket dataset, we examined the record at REMARK 900 in their PDB files where a list of related PDB entries are provided, and found apo proteins for 96 proteins. Proteins were removed if there are less than 5 holo proteins in the database.

To define a ligand binding pocket of an apo protein, residues are determined as ligand-binding if their corresponding residues in the holo proteins are in contact with the ligand. The database search was performed in the same way as it was done for holo proteins: A binding pocket of an apo protein was compared against the pockets in the non-redundant binding pocket database and ligands were predicted for the query apo pocket by computing the Pocket-Score_w from the ranked list of retrieved pockets.

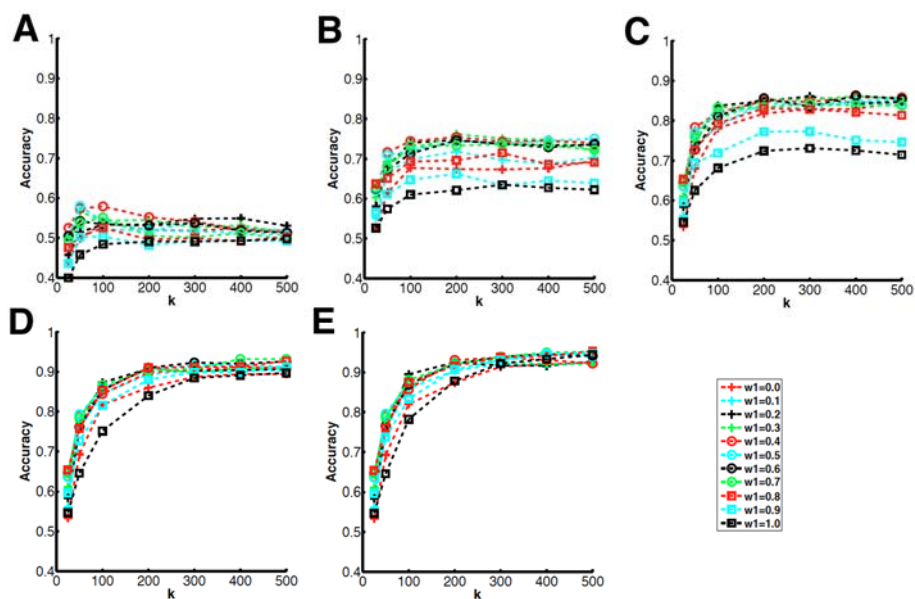
Table S4. An example of top 10 hits by a query pocket, 1gco_A that binds NAD

Rank	PDB ID of the pocket	Binding ligand (PDB code)	Structure of ligand	SIMCOMP Score ^b
Query	1gco_A	NAD		1.0
1	1lj8_A	NAD		1.0
2	1ebw_AB	BEI		0.17
3	3b4y_A	F42_FLC ^a		0.26
4	3oa2_ACD	NAD		1.0

5	1bxk_A	NAD		1.0
6	3c1o_A	NAP		0.88
7	1nuq_A	DND		0.87
8	2jhf_A	NAD		1.0
9	1nyt_B	NAP		0.88
10	2whh_A	PPN_GLU_PPN_GLU		0.11

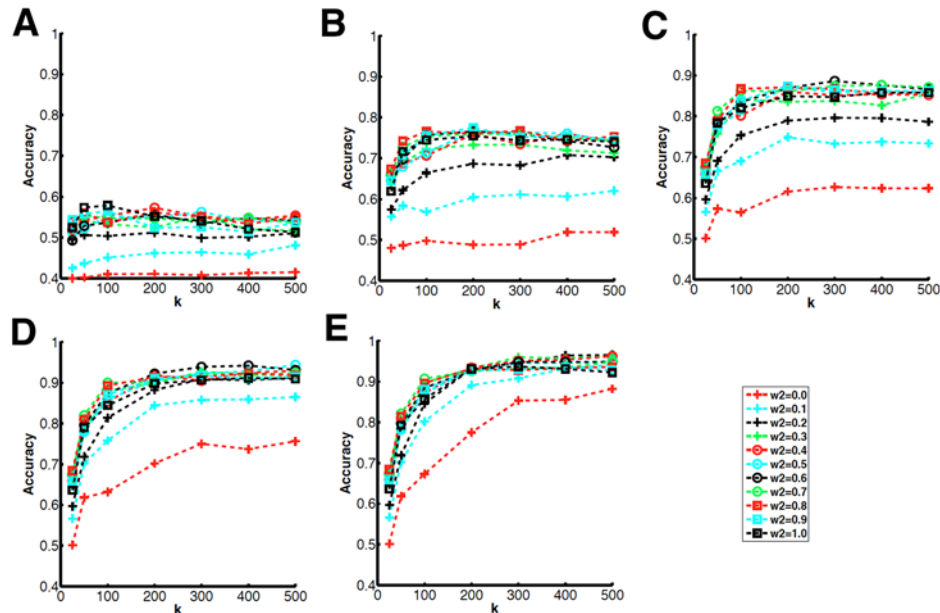
- a) Two ligands, F42 and FLC bind to the pocket and they are closer than 4.5 Å.
b) SIMCOMP score to NAD, the ligand that is binding to the query pocket.

Figure S1. Average accuracy using *M*Score with different combinations of *k* and *w*1 values.



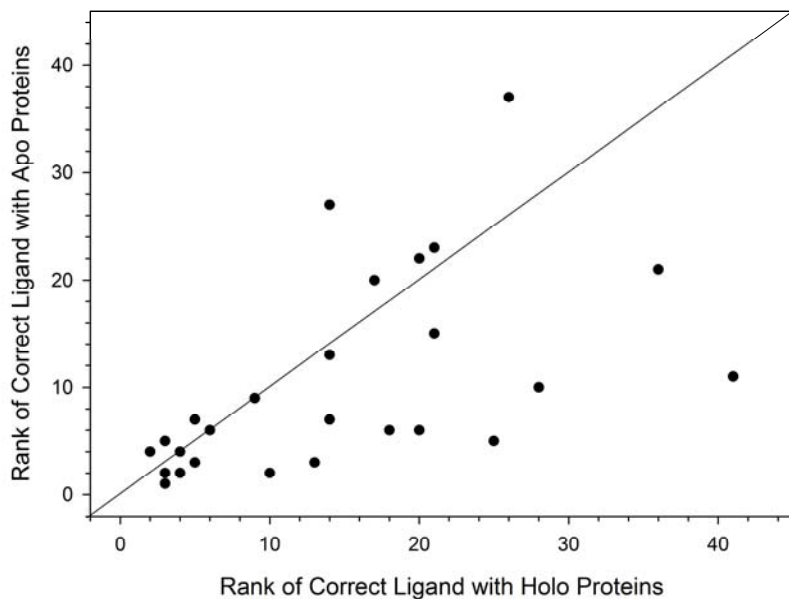
Different lines correspond to results with different *w*1. **A**, Top5 accuracy; **B**, Top10; **C**, Top15; **D**, Top20; and **E**, Top25 accuracy. The panels for Top 10 and Top 15 accuracy are also shown in Figure 2 in the main text.

Figure S2. Average accuracy using *TScore* with different combinations of k and w_2 values. w_1 is set to 0.4.



A, Top5 accuracy; B, Top10; C, Top15; D, Top20; and E, Top25 accuracy.

Figure S3. Ranks of the correct ligands for holo and apo proteins.



Binding ligands are predicted for holo and apo proteins in Table S3 and the ranks of the correct ligands are plotted. Holo and apo protein pairs with – in the rank column were excluded from this plot.