Supplemental Information for article entitled "Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions"

Gang Hu[1], Jianzhao Gao[1], Kui Wang[1*], Marcin J. Mizianty[2], Jishou Ruan[1,3] and Lukasz Kurgan[2*]

[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, People's Republic of China 300071

[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada T6G 2V4

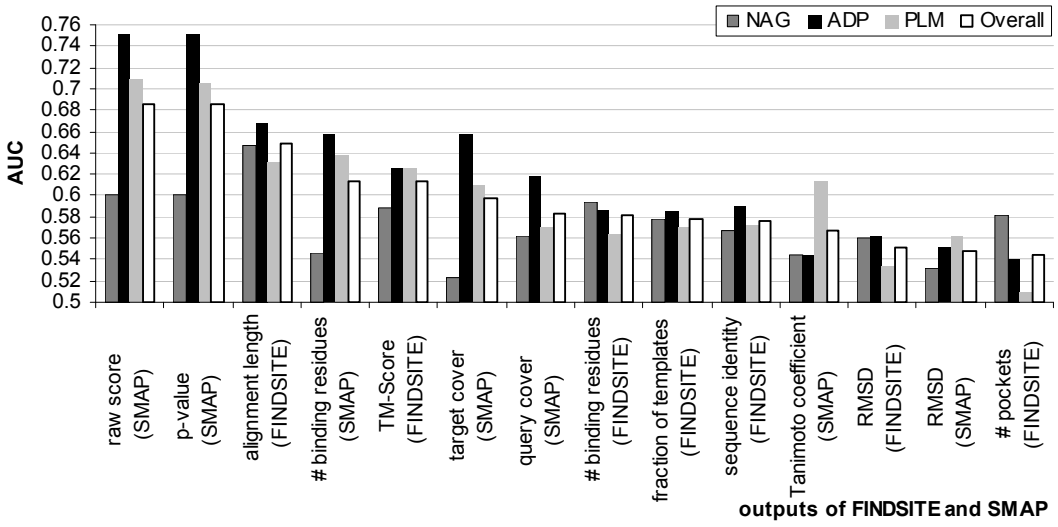
[3] State Key Laboratory for Medicinal Chemical Biology, Nankai University, Tianjin, People's Republic of China 300071

*** Corresponding authors:**

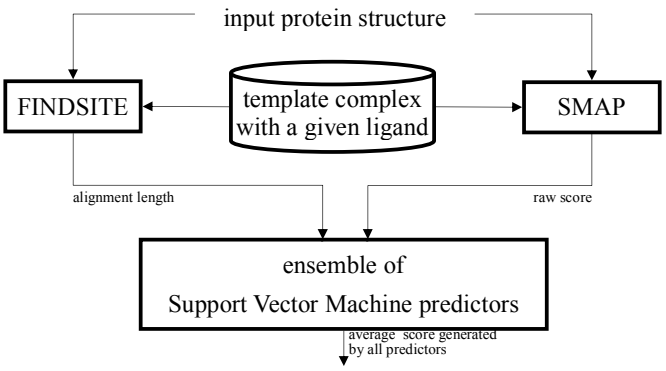
L. Kurgan: lkurgan@ece.ualberta.ca, phone 780-492-5488, fax 780-492-1811

K. Wang: wangkui@nankai.edu.cn

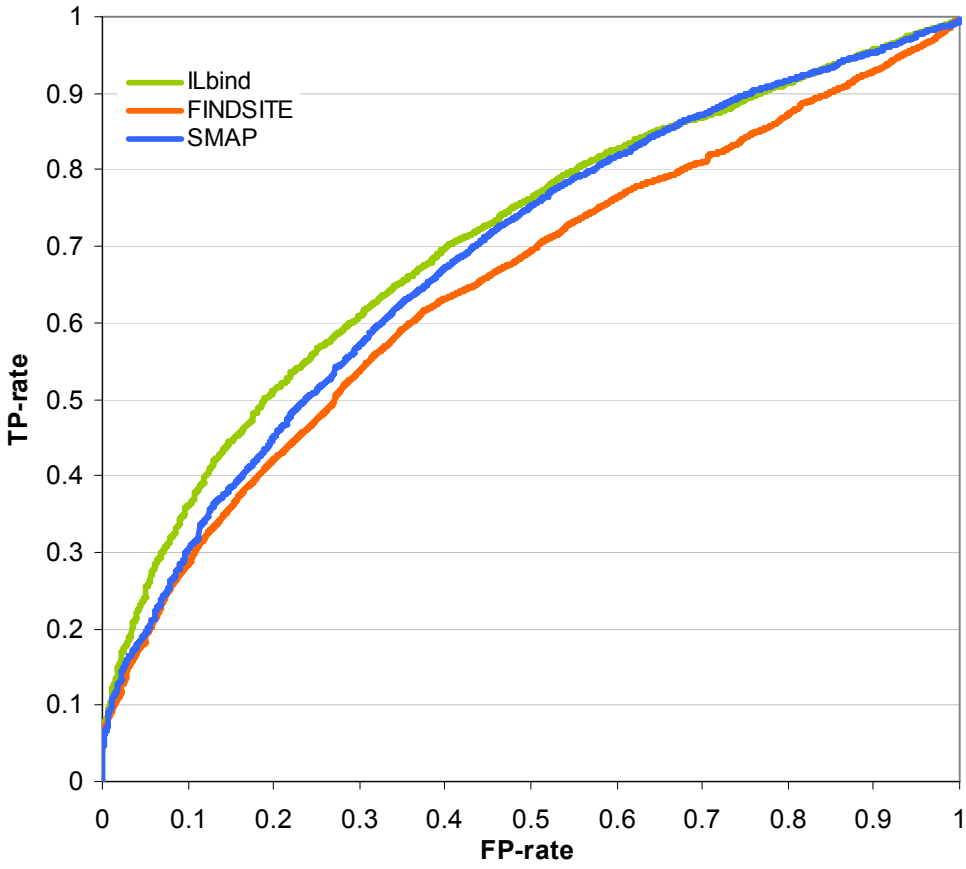
Supplemental Figures



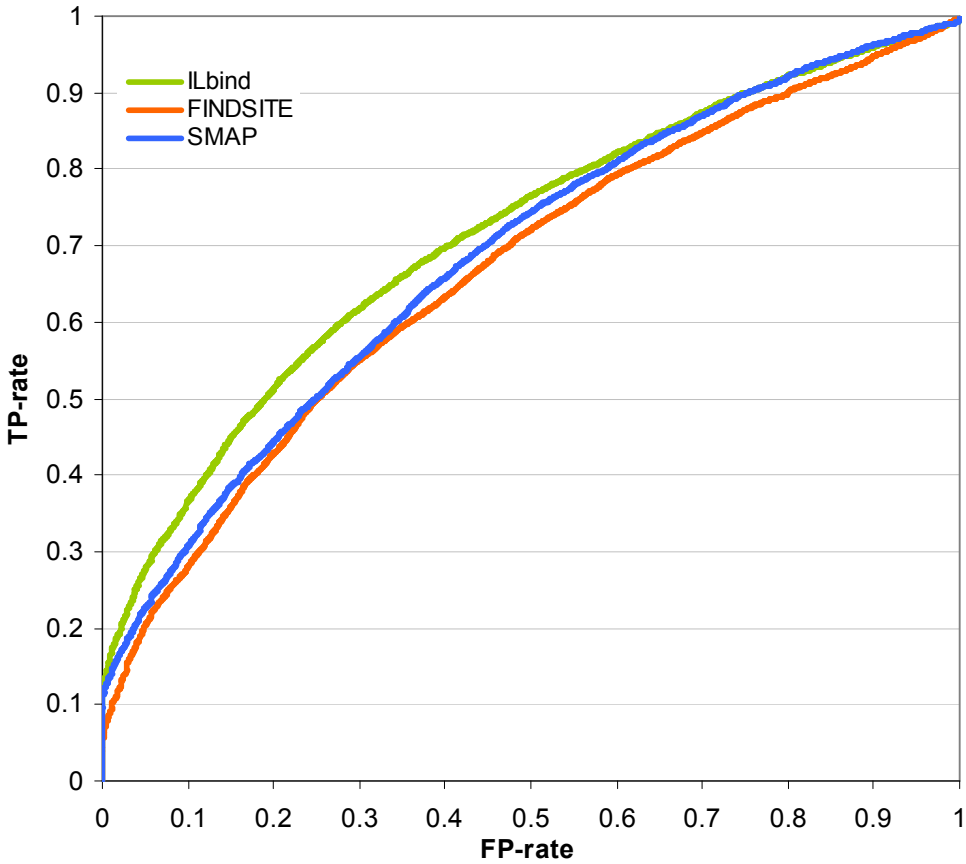
Supplementary Figure S1. Comparison of average (over 5 diverse templates) AUC values that quantify quality of the prediction of binding proteins for the three selected ligands (NAG, ADP and PLM) including the average over the three ligands (overall); related to Table 1 and Figure 1. The predictions were performed on the redundant benchmark datasets using each of the 14 outputs generated by FINDSITE and SMAP, which are shown on the x-axis. The outputs are sorted in the descending order (left to right) by their overall AUCs.


input protein structure
FINDSITE
template complex
with a given ligand
SMAP
alignment length
raw score
ensemble of
Support Vector Machine predictors
average score generated
by all predictors
predicted propensity of binding to the given ligand (score)

A


ILbind
FINDSITE
SMAP
TP-rate
FP-rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

B



C

Supplementary Figure S2. Detailed results for predictions with FINDSITE, SMAP, and ILbind and architecture of the ILbind predictor; related to Table 2.

Panel S2A is an overview of the architecture of the consensus-based inverse ligand binding predictor ILbind.

Panel S2B shows the averaged ROC curves for FINDSITE, SMAP and the consensus based ILbind method calculated over the 5 templates and the 3 selected ligands: NAG, ADP and PLM.

Panel S2C gives the averaged ROC curves for FINDSITE, SMAP and the consensus based ILbind method calculated over the independent set of 35 ligands.

Supplemental Tables

Supplementary Table S1. Detailed results for predictions with FINDSITE, SMAP, and ILbind; related to Table 2.

Sub-table S1A gives the AUCs for the five templates and the three selected ligands: NAG, ADP and PLM. The last column gives AUCs for an SVM-based consensus that uses all 14 features (all combined outputs of FINDSITE and SMAP).

Sub-table S1B provides the AUCs for the test on the independent set of 35 ligands. The first column is ligands name and its corresponding number of non-hydrogen atoms, which is used in sub-table S1C. The next three columns give AUCs for FINDSITE, SMAP and the consensus-based ILbind predictor that was build using the three selected ligands (NAG, ADP, and PLM) and tested on these 35 different ligands. The last row shows the average AUCs over the 35 ligands. The best results among the three predictors for each ligand are shown in bold font.

Sub-table S1C shows AUCs and significance of differences in AUCs for varying ligand sizes based on results on the independent set of 35 ligands. The ligands are grouped based on the number of their non-hydrogen atoms *n*. The last three columns show the *p*-values that quantify significance of differences between FINDSITE and SMAP (F vs. S), FINDSITE and ILbind (F vs ILb) and SMAP and ILbind (S vs ILb). Best results for each ligand size bin are shown in bold font.

S1A

Ligand	Template	FINDSITE	SMAP	ILbind	All 14 features
NAG	1zag	0.75	0.61	0.75	0.66
	1nql	0.70	0.59	0.67	0.62
	2ciy	0.61	0.57	0.61	0.61
	2wfo	0.55	0.56	0.57	0.61
	3c45	0.62	0.67	0.67	0.67
	Average	0.646	0.600	0.654	0.634
ADP	1gzf	0.63	0.70	0.71	0.73
	3c9u	0.66	0.70	0.73	0.74
	1cqi	0.70	0.76	0.76	0.71
	2zpa	0.69	0.79	0.79	0.60
	3cnz	0.66	0.80	0.79	0.75
	Average	0.668	0.750	0.756	0.706
PLM	2iu8	0.69	0.61	0.66	0.66
	3lsj	0.59	0.75	0.74	0.69
	2ies	0.59	0.69	0.70	0.68
	3fys	0.60	0.71	0.69	0.62
	2g87	0.69	0.78	0.78	0.76
	Average	0.632	0.708	0.714	0.682

Ligand name	# atoms	FINDSITE	SMAP	ILbind
ATP	31	0.79	0.81	**0.84**
ANP	31	0.74	0.85	**0.88**
APC	31	0.66	0.69	**0.70**
UDP	25	0.74	0.73	**0.77**
CIT	13	0.57	0.58	**0.59**
BOG	20	0.61	0.62	**0.65**
TRP	15	0.75	**0.87**	0.86
ARG	12	**0.79**	0.54	0.69
NAP	48	0.60	**0.71**	0.70
BTB	14	**0.66**	0.57	0.60
PG4	13	0.61	0.59	**0.63**
EPE	15	**0.55**	0.53	0.54
AMP	23	0.69	0.64	**0.70**
MES	12	**0.65**	0.56	0.63
COA	48	0.66	**0.70**	**0.70**
GDP	28	0.54	**0.62**	0.61
1PE	13	0.66	0.62	**0.69**
HEC	43	0.85	**0.97**	**0.97**
FLC	13	0.60	0.59	**0.61**
ACO	51	0.53	**0.67**	0.63
NDP	48	0.62	**0.72**	0.71
GTP	32	0.73	**0.80**	**0.80**
2PE	28	**0.59**	0.53	0.57
SUC	23	0.69	0.61	**0.70**
SAM	27	0.68	**0.83**	0.81
MAN	12	0.68	0.60	**0.69**
FMN	31	0.54	**0.56**	**0.56**
BGC	12	**0.77**	0.64	0.75
ADN	19	0.65	**0.66**	0.65
ACP	31	0.71	**0.82**	0.80
SAH	26	0.58	0.62	**0.63**
P6G	19	**0.76**	0.68	**0.76**
FAD	53	0.72	0.69	**0.74**
NAD	44	0.77	0.91	**0.92**
HEM	43	0.56	**0.84**	0.83
Average		0.666	0.685	**0.713**

S1B

	Average AUCs			P values		
Ligand size	**FINDSITE**	**SMAP**	**ILbind**	**F vs. S**	**F vs. ILb**	**S vs. ILb**
$n \leq 14$	**0.67**	0.59	0.65	0.02	0.42	<0.01
$15 \leq n \leq 26$	0.67	0.66	**0.70**	0.76	0.06	0.03
$27 \leq n \leq 31$	0.66	0.71	**0.72**	0.05	0.01	0.40
$n > 31$	0.67	**0.78**	**0.78**	0.01	<0.01	0.89
all 35 ligands	0.67	0.68	**0.71**	0.27	<0.01	<0.01

S1C

Supplemental Experimental Procedures

Selection of Representative Ligands

A detailed, step-by-step description of the ligand selection process:

1. *Extract all protein-ligand complexes from PDB*. We collect all protein-ligand complexes from PDB in April 2011, which include total of 14,286 ligands.

2. *Select biologically relevant ligands*. Following (Dessailly et al., 2008), we remove the metal ions, lipids and peptides, ligands with less than 10 heavy atoms, and ligands in complexes with contact numbers less than 70. Finally, 145 ligands are left after we remove the ligands for which the numbers of binding proteins is <10.

3. *Reduction of sequence similarity***.** We cluster protein sequences for each ligand using Blastclust with 25% similarity using one chain, with the maximal contact number with the ligand, for each target protein. We pick one protein with the maximal contact number per cluster, which means that the remaining proteins have pairwise similarity below 25%. There are 78 ligands left after we remove the ligands with the number of the remaining proteins <10. The number of target proteins ranges between 239 and 10; 8 ligands have over 100 partner proteins.

4. *Reduction of structure similarity***.** We align binding proteins for each remaining ligands using fr-Tm-Align (Pandit and Skolnick, 2008). We reduce the binding proteins to a set in which the pairwise structural similarity is below 0.4 and we remove the ligands with the number of the remaining proteins <10. As a result, 38 ligands with the corresponding number of proteins that ranges between 10 and 59 are left.

5. *Ligand Clustering***.** We cluster the remaining 38 ligands in two steps. First, we calculate fingerprint (structural descriptor) for each ligand using the rcdk package in R. Next, we cluster these fingerprints using k-mean algorithms with the Tanimoto distance into 3 clusters. Cluster centers are represented using mode value (the most frequent binary value) across the corresponding dimensions in the fingerprints from all ligands in a given cluster. The centers are initialized as the fingerprints of 3 ligands that bind the largest number of proteins calculated in step 4. The clustering has converged after 4 iterations and we selected the ligand with the largest number of binding proteins to represent each of the 3 resulting clusters. As a result, NAG, ADP and PLM ligands were selected.

Supplemental References

1. Dessailly, B.H., Lensink, M.F., Orengo, C.A., and Wodak, S.J. (2008). LigASite-a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. *36*, D667-673

2. Pandit, S.B. and Skolnick, J. (2008). Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. BMC Bioinformatics *9*, 531