Supporting Information

Prediction of binding free energy of protein-ligand complexes with a hybrid MM/GBSA and machine learning method

Lina Dong,†‡ Xiaoyang Qu,† Yuan Zhao,¶ Binju Wang†*

†State Key Laboratory of Physical Chemistry of Solid Surfaces and Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 360015, P. R. China.

‡Collaborative Innovation Center of Chemistry for Energy Materials, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, China

¶The Key Laboratory of Natural Medicine and Immuno-Engineering, Henan University, Kaifeng, China

*Correspondence to: wangbinju2018@xmu.edu.cn

Table S1. Brief description and the tuned hyperparameters of each machine learning method.

ML method	Brief description	Tuned hyperparameters
Ridge regression[1] (RR)	Ridge regression is a kind of biased estimation regression method for collinear data analysis. In essence, it is an improved least square estimation method. By abandoning the unbiased property of least square method, the regression coefficient is obtained at the cost of losing part of information and reducing precision.	alpha=0.1
Decision tree regression[2] （DT）	Decision tree is a simple and commonly used supervised learning model based on tree. According to the use, it can be divided into classification tree and regression tree. A regression tree corresponds to a partition of the input space (that is, the feature space) and the output value on the partition unit.	max_depth=3.0
Extra trees regression[3] （ET）	Extreme tree regression is also a model based on tree regression, but the major discrepancy is that it uses extra-trees to replace the conventional decision-trees. Unlike conventional top-down tree construction algorithms, it applies randomization in both selecting features and cut-point features, thus leading to its totally randomized trees in an extreme case.	n_estimators=50

SVM regression[4] （SVM）	SVM regression is a variation of support vector machine (SVM), and performs the regression task by searching a hyperplane with the optimized sum of the distances from the data points to the hyperplane.	max_iter=3000
Random forest regression[5] （RF）	Random forest is an ensemble learning approach that can effectively improve the predictive accuracy by combining the idea of bagging and random selection of features to establish a decision tree set of control variables. It is essentially an improvement of decision tree algorithm.	n_estimators=[100,200,300,400,500,600] max_depth=[3,5,7,9]
Neural network regression[6] （DNN）	Neural network is one of the typical deep learning methods which include input layers, hidden layers and output layers. For the regression problem in supervised learning, it optimizes the weight and bias between layers according to the back propagation of the loss between the target value and the predicted value.	max_iter=500 hidden_layer_sizes=1000
Extreme gradient boosting[7] (XGB)	Extreme gradient boosting is essentially a kind of gradient boosting decision tree(GBDT), which is used widely by data scientists to achieve state-of-the-art results on many machine learning challenges. It improves the efficiency of calculation loss and optimization.	n_estimators=[100,200,300,400,500,600] max_depth=[3,5,7,9]

Table S2. Performance (Pearson's correlation coefficient and MSE value) of seven machine learning models on the validation set with three different feature set.

Features	Model name	MSE	Pearsonr
MM/GBSA（G）	Linear regression（LR）	4.97	0.503
	Ridge regression（RR）	4.97	0.503
	Decision tree regression（DT）	4.63	0.548
	Extra trees regression（ET）	5.91	0.511
	SVM regression（SVM）	5.08	0.507
	Random forest regression（RF）	4.69	0.551
	Neural network regression（DNN）	23.70	0.508
	Extreme gradient boosting (XGB)	6.34	0.501
MM/GBSA+X-Score（G+X）	Linear regression（LR）	4.93	0.538
	Ridge regression（RR）	4.93	0.538
	Decision tree regression（DT）	5.05	0.522
	Extra trees regression（ET）	4.26	0.598
	SVM regression（SVM）	5.21	0.538
	Random forest regression（RF）	4.12	0.616
	Neural network regression（DNN）	21.07	0.559

	Extreme gradient boosting (XGB)	6.28	0.550
MM/GBSA+X-Score+ Ligand based （G+X+L）	Linear regression（LR）	4.25	0.610
	Ridge regression（RR）	4.25	0.610
	Decision tree regression（DT）	4.66	0.547
	Extra trees regression（ET）	3.65	0.656
	SVM regression（SVM）	4.44	0.608
	Random forest regression（RF）	3.73	0.647
	Neural network regression（DNN）	11.93	0.596
	Extreme gradient boosting (XGB)	4.96	0.598

Table S3. Scoring and docking power of GXLE and scoring functions benchmarked in CASF-2016.

method	scoring power	ranking power	method	scoring power	ranking power
GXLE	0.762	0.63	PLP2@DS	0.563	0.589
deltaVinaRF20	0.732	0.626	Affinity-dG@MOE	0.552	0.604
X-Score	0.631	0.604	LigScore2@DS	0.54	0.608
X-ScoreHS	0.629	0.547	D-Score@SYBYL	0.531	0.577
deltaSAS	0.625	0.588	LUDI2@DS	0.526	0.629
X-ScoreHP	0.621	0.573	GlideScore-SP	0.513	0.419
ASP@GOLD	0.617	0.553	LUDI3@DS	0.502	0.532
ChemPLP@GOLD	0.614	0.633	GBVI_WSA-dG@MOE	0.496	0.489
X-ScoreHM	0.609	0.603	LUDI1@DS	0.494	0.612
AutodockVina	0.604	0.528	GlideScore-XP	0.467	0.257
DrugScore2018	0.602	0.607	Jain@DS	0.457	0.521
DrugScoreCSD	0.596	0.63	LigScore1@DS	0.425	0.599
ASE@MOE	0.591	0.439	PMF@DS	0.422	0.537
ChemScore@SYBYL	0.59	0.593	GoldScore@GOLD	0.416	0.284
PLP1@DS	0.581	0.582	London-dG@MOE	0.405	0.593
ChemScore@GOLD	0.574	0.526	PMF@SYBYL	0.262	0.449

G-Score@SYBYL	0.572	0.591	PMF04@DS	0.212	0.481
Alpha-HB@MOE	0.569	0.535			

The scoring and ranking power of scoring functions benchmarked in CASF-2016 was obtained from Ref[8].

Table S4. PDBid of 10 kind of selected targets from PDBbind general set.

target	PDBid							
BACE-1	1llb	1ymx	2b8l	2b8v	2f3e	2f3f	2hiz	2hm1
	2iqg	2irz	2is0	2oah	2ohm	2ohp	2ohq	2ohr
	2ohs	2oht	2ohu	2p83	2p8h	2q9n	2qk5	2qmd
	2qmf	2qp8	2qzk	2qzl	2va5	2va6	2va7	2vie
	2viz	2vj6	2vj7	2vj9	2vnm	2vnn	2wez	2wf0
	2wf1	2wf2	2wf3	2wf4	2xfi	2xfj	2xfk	3dv1
	3dv5	3fkt	3hvg	3hw1	3in4	3ine	3inf	3ivh
	3ivi	3k5c	3k5d	3k5g	3kn0	3l58	3l5c	3l5d
	3l5e	3l5f	3lhg	3lnk	3lpj	3msj	3msk	3msl
	3s2o							
CHK1	2ayp	2cgw	2cgx	2e9n	2e9o	2gdo	2ghg	2r0u
	2ywp	3f9n	3ot3	3ot8	3pa3	3pa4	3pa5	
DPP4	1rwq	2bub	2fjp	2ogz	2onc	2rip	3ccb	3g0b
	3g0d	3g0g	3kwf	3kwj	3g0c			
ER	1sj0	1uom	1x7r	1xp1	1xp6	1xpc	2b1z	
LTA-4H	3cho	3chp	3chq	3chr	3chs	3fh5	3fh8	3fhe
	3fts	3ftu	3ftv	3fty	3fu0	3fu3	3fu6	3fue
	3fuf	3fuh	3fuj	3fuk	3ful	3fum		
P38a	1a9u	1bl7	1bmk	1di9	1w84	1ywr	1zyj	1zzl
	2gfs	2qd9	3ds6	3dt1	3gfe	3itz	3mvm	3nnw

	3nww	3ocg						
PPAR	1nyx	1zeo	2f4b	2g0h	3bc5	3fei	3g8i	3g9e
	3gwx	3gz9	3kdt					
PTP1B	2cm7	2cm8	2cmb	2cmc	2cng	2cnh	2cni	2f6v
	2f71	2veu	2vev	2vew	2vex	2vey		
Thrombin	1riw	1way	1wbg	2a2x	2ank	2anm	2bdy	2c8w
	2c8x	2c8y	2c90	2c93	2feq	2fes	2pks	
Renin	1bil	1bim	1hrn	1rne	2bks	2bkt	2v0z	2v10
	2v11	2v12	2v13	2v16	3g6z	3g70	3g72	3gw5
	3km4	3oad	3oag	3oot	3oqf	3oqk		

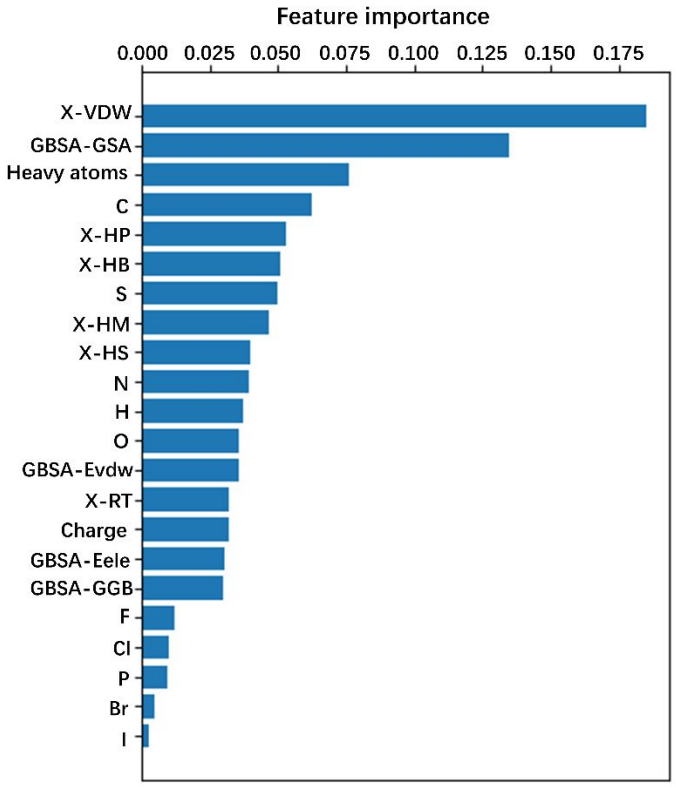

Figure S1. Feature importance. Feature importance are calculated based on the number of times a feature is used to split the data across all trees. Here, all 22 features are shown.

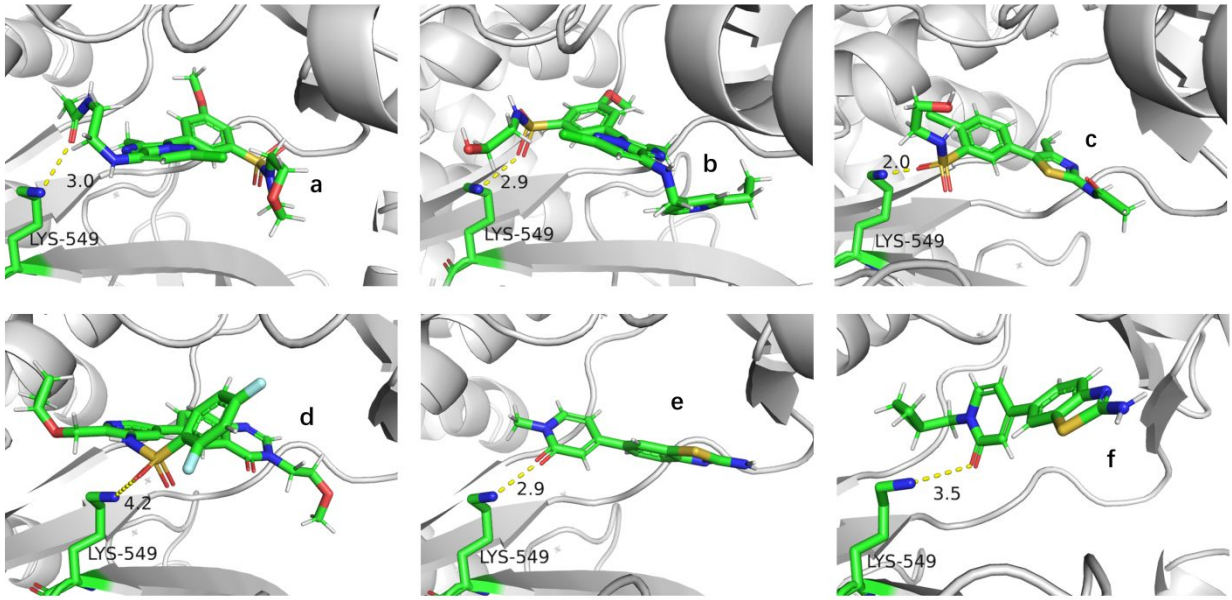

Figure S2. Figure in the upper right corner shows the crystal structure (PDB id: 4D0L) of PI4KIIIβ in complex with N-(5-(4-chloro-3-(2-hydroxy-ethylsulfamoyl)-phenylthiazole-2-yl)-acetamide (named **c**), which was used as the receptor structure in molecular docking. And five selected PI4KIIIβ inhibitors (named **a, b, d, e, f**) used for the evaluation of ranking power of GXLE together with their binding modes predicted by molecular docking.

REFERENCES

(1) P. J. Brown, J. V. Z., Adaptive Multivariate Adriate Ridge Regression. *The Annals of Statistics* **1980**, 8, 64-74.

(2) Brodley, C. E.; Utgoff, P. E., MULTIVARIATE DECISION TREES. *Mach. Learn.* **1995**, 19, 45-77.

(3) Geurts, P.; Ernst, D.; Wehenkel, L., Extremely randomized trees. *Mach. Learn.* **2006**, 63, 3-42.

(4) Brereton, R. G.; Lloyd, G. R., Support vector machines for classification and regression. *Analyst* **2010**, 135, 230-267.

(5) Breiman, L., Random forests. *Mach. Learn.* **2001**, 45, 5-32.

(6) Sainath, T. N.; Vinyals, O.; Senior, A.; Sak, H.; Ieee CONVOLUTIONAL, LONG SHORT-TERM MEMORY, FULLY CONNECTED DEEP NEURAL NETWORKS. In 2015 Ieee International Conference on Acoustics, Speech, and Signal Processing; Ieee: New York, 2015, pp 4580-4584.

(7) Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Krishnapuram, B., Shah, A., Aggarwal, C., Shen, D., Rastogi, R.,Eds.; ACM: San Francisco, CA, 2016, pp 785-794.

(8) Su, M.; Yang, Q.; Du, Y.; Feng, G.; Liu, Z.; Li, Y.; Wang, R., Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* **2019**, 59, 895-913.