WDL-RF: Predicting Bioactivities of Ligand Molecules Acting with G Protein-coupled Receptors by Combining Weighted Deep Learning and Random Forest

Jiansheng Wu, Qiuming Zhang, Weijian Wu, Tao Pang, Haifeng Hu, Wallace K.B. Chan, Xiaoyan Ke, Yang Zhang

Supporting Information

Table S1. List of classes and subfamilies of GPCR datasets used in this study.

Class	**Subfamilies**	**GPCR ID (# of Known Ligands)**
A	Aminergic receptors	P08908(2294), P50406(1421), P08912(369), P35348(1027), P21917(1679), Q9Y5N1(2092),
	Peptide receptors	P30968(1124), P24530(1019), Q99705(2052), P35372(3828), P46663(452), P35346(689), P21452(696),
	Nucleotide receptors	P30542 (3016),
	Lipid receptors	Q99500(317), Q9Y5Y4(641), P34995(236),
	Protein receptors	P51677(781),
	Melatonin receptors	P48039(684)
	Steroid receptors	Q8TDU6(1153)
	Alicarboxylic acid receptors	Q8TDS4(271)
	Sensory receptors	no known ligands for all of these GPCRs
	Orphan receptors	Q9HC97(1589)
B	Peptide receptors	P47871(1129)
	Adhesion receptors	no known ligands for all of these GPCRs
C	Ion receptors	P41180(940)
	Amino acid receptors	Q14416(1810)
	Sensory receptors	no known ligands for all of these GPCRs
	Orphan receptors	no known ligands for all of these GPCRs
F	Protein subfamily	Q99835(1523)
Taste 2	Sensory receptors	no known ligands for all of these GPCRs
Other	Orphan receptors	no known ligands for all of these GPCRs

Table S2. Performance of Molecular Fingerprints from different module units.

EC[a]	Layer[b]	GPCR Datasets						
		P30968	P24530	Q99705	P35372	P30542	P08908	P50406
RMSE (↓)	WDL	**1.21**	**1.04**	**1.33**	**1.51**	**0.96**	**1.50**	1.45
	1st	2.02*	1.95*	2.22*	2.35*	1.95*	2.31*	2.52*
	2nd	1.99*	1.81*	2.36*	2.21*	1.90*	2.17*	2.57*
	3rd	1.75*	1.79*	2.12*	2.13*	1.81*	2.07*	2.48*
	4th	1.31*	1.24*	1.50*	1.60	0.98	1.55	**1.42**
	5th	1.59*	1.73*	2.03*	1.98*	1.69*	1.97*	2.41*
	6th	1.61*	1.71*	1.99*	2.01*	1.69*	2.01*	2.39*
r^2 (↑)	WDL	**0.86**	**0.85**	**0.82**	**0.74**	**0.86**	**0.70**	0.79
	1st	0.58*	0.44*	0.49*	0.36*	0.41*	0.27*	0.38*
	2nd	0.60*	0.52*	0.41*	0.44*	0.44*	0.37*	0.36*
	3rd	0.70*	0.55*	0.54*	0.48*	0.50*	0.47*	0.40*
	4th	0.84	0.77*	0.77	0.71	0.85	0.67	**0.80**
	5th	0.76*	0.59*	0.57*	0.51*	0.55*	0.53*	0.42*
	6th	0.74*	0.61*	0.59*	0.51*	0.53*	0.51*	0.45*
q^2 (↑)	WDL	**0.85**	**0.81**	**0.81**	**0.73**	**0.85**	**0.69**	**0.79**
	1st	0.58*	0.43*	0.48*	0.34*	0.40*	0.27*	0.37*
	2nd	0.60*	0.50*	0.40*	0.42*	0.43*	0.34*	0.34*
	3rd	0.69*	0.51*	0.52*	0.46*	0.48*	0.41*	0.39*
	4th	0.82	0.75*	0.76	0.70	0.84	0.67	**0.79**
	5th	0.75*	0.57*	0.55*	0.50*	0.53*	0.54*	0.41*
	6th	0.72*	0.56*	0.58*	0.49*	0.50*	0.53*	0.42*

[a]Evaluation Criterion: ↑ (↓) indicates the larger (smaller), the better the model performance; the best results on each evaluation criterion are highlighted in boldface.

[b]Layers: The WDL is the default molecular fingerprint with four module units used in this paper, i.e., the weighted molecular fingerprint of all module units. The *1st* to *6th* respectively denote the molecular fingerprint generated by the first to the sixth module unit of our approach WDL-RF.

GPCR datasets: '*' indicates the performance of the molecular fingerprint from a module unit is significantly worse than that of WDL based on Wilcoxon signed-ranked test.