Electronic Supplementary Material (ESI) for Digital Discovery.

MPSM-DTI: Prediction of Drug-Target Interaction via Machine Learning based on Chemical Structure and Protein Sequence

Yayuan Peng, Jiye Wang, Zengrui Wu*, Lulu Zheng, Biting Wang, Guixia Liu, Weihua Li, Yun Tang*

Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China.

*Corresponding authors. E-mail: ytang234@ecust.edu.cn (Y.T.); zengruiwu@ecust.edu.cn (Z.W.).

Model construction

Five machine learning methods were used to build models for DTI prediction, including decision tree (DT), bagging, gradient boost decision tree (GBDT), k-nearest neighbors (KNN), and support vector machine (SVM).

DT. By a set of if-then-else decision rules learned from data features, DT created a tree model to solve very difficult problems [1]. The deeper the tree is, the more complex the DT model is.

Bagging. Bagging trains a number of base estimators parallelly each from a different *bootstrap sample* by calling a base learning algorithm [2]. Here the base estimators are DT. After obtaining the base estimators, bagging combines all individual predictions and then forms a final prediction by majority voting.

GBDT. GBDT is a kind of boosting ensemble strategy with training DT base estimator successively [3]. The main idea of GBDT is to find DT hypothesis $h_t(X)$ to make the loss function lowest in every training process.

KNN. KNN is to assign the classification of unclassified sample points depending on a number of closest previously classified sample points [4]. The distance can be any metric measure, among which standard Euclidean distance is the most common choice. In scikit-learn, the contribution of neighbor samples can be adjusted according to far or near distance by altering the parameter of *weights*.

SVM. SVM is a class of supervised machine learning methods for classification, regression and outliers detection [5]. SVC (support vector classification) is the classification algorithm for two class or multiclass classification. The main idea of SVC is to find a hyperplane which could categorize difficult samples by mapping data into high dimensional space [6]. Kernel tricks was critical when mapping data from low dimensional to high dimensional space. Usually the *rbf* kernel was used.

Table S1. The predicted results of 100 active compound target interaction pairs using MPSM-DTI model, SwissTarget, NetInfer and ChemMapper webserver. The "1" means true prediction and "0" means false prediction. The results of SDTNBI (Top 20) and bSDTNBI (Top 20) were not list below for the limited space.

No.	Target	Compound ID	MPSM-DTI	SwissTarget	SDTNBI (Top 50)	bSDTNBI (Top 50)	ChemMapper	Ref.
1	DHCR7	C-01	1	0	1	1	0	7
2		C-02	1	1	0	0	0	8
3		C-03	1	1	0	0	0	8
4		C-04	1	1	0	0	0	8
5		C-05	1	1	0	0	0	8
6		C-06	1	0	1	1	0	9
7		C-07	1	0	0	0	0	9
8		C-08	1	0	0	0	0	9
9	HTR1F	C-09	1	1	1	1	1	10
10		C-10	1	1	0	0	0	11
11		C-11	1	1	0	0	0	11
12		C-12	1	1	0	0	0	12
13		C-13	1	1	0	0	0	12
14		C-14	1	1	0	0	0	12
15		C-15	1	1	0	0	0	12
16		C-16	1	1	0	0	0	12
17		C-17	1	1	0	0	0	12
18		C-18	1	1	0	0	0	12
19		C-19	1	1	0	0	0	12
20		C-20	1	1	0	0	1	12
21		C-21	1	1	1	1	1	12
22		C-22	1	1	0	0	0	13
23	LTB4R	C-23	1	1	1	1	0	14
24		C-24	1	1	1	1	1	15
25		C-25	1	1	1	1	1	15
26		C-26	1	1	0	1	1	15
27		C-27	1	0	1	1	1	15
28		C-28	1	0	1	1	1	15
29		C-29	1	1	1	1	1	15
30		C-30	1	0	0	1	1	15
31		C-31	1	0	1	1	1	15
32		C-32	1	1	1	1	1	15
33		C-33	1	0	1	1	1	15
34		C-34	1	0	1	1	1	15
35		C-35	1	0	1	1	1	15
36		C-36	1	0	1	1	1	15
37	CYSLTR2	C-37	1	1	0	0	0	16
38		C-38	0	1	0	0	1	16

39		C-39	1	1	1	1	0	[16]
40		C-40	1	1	0	1	1	[16]
41		C-41	1	1	1	1	1	[16]
42		C-42	0	1	0	0	1	[17]
43		C-43	0	1	0	0	1	[17]
44		C-44	0	0	0	0	0	[17]
45		C-45	1	0	1	1	0	[17]
46		C-46	1	1	0	0	0	[16]
47		C-47	1	1	0	0	1	[16]
48		C-48	1	1	0	0	1	[16]
49		C-49	1	1	0	0	1	[16]
50		C-50	1	1	0	0	0	[16]
51		C-51	1	1	0	0	0	[16]
52	GRIK3	C-52	1	1	1	1	0	[18]
53		C-53	1	1	1	1	1	[19]
54		C-54	1	1	1	1	1	[19]
55		C-55	0	1	0	1	0	[19]
56		C-56	0	1	0	1	0	[19]
57		C-57	1	1	0	1	0	[19]
58		C-58	1	0	1	1	0	[20]
59		C-59	1	1	0	0	0	[21]
60	GPER1	C-60	1	1	1	1	0	[22]
61		C-61	1	1	1	1	0	[22-24]
62		C-62	1	0	0	0	0	[25]
63		C-63	1	1	0	0	0	[26]
64		C-64	1	1	1	1	0	[27]
65		C-65	1	0	0	0	0	[26]
66		C-66	1	1	0	0	0	[26]
67		C-67	1	0	0	0	0	[26]
68		C-68	1	1	1	1	0	[24]
69		C-69	1	1	1	1	0	[23]
70	PTGIR	C-70	1	0	0	1	1	[28]
71		C-71	1	1	0	1	1	[28]
72		C-72	1	0	0	1	1	[28]
73		C-73	1	0	0	1	1	[28]
74		C-74	1	0	0	1	1	[28]
75		C-75	1	1	1	1	0	[28]
76		C-76	1	0	1	1	0	[28]
77		C-77	1	0	1	1	0	[28]
78		C-78	1	0	1	1	0	[28]
79		C-79	1	0	0	1	0	[28]
80		C-80	1	0	1	1	1	[28]
81		C-81	1	0	0	1	0	[28]
82		C-82	1	1	1	1	0	[28]

83		C-83	1	0	0	0	1	[28]
84		C-84	1	0	0	1	0	[28]
85		C-85	1	0	0	1	0	[28]
86		C-86	1	1	1	1	0	[29]
87		C-87	1	1	1	1	0	[30]
88		C-88	1	1	0	0	0	[31]
89		C-89	1	0	0	0	0	[32]
90		C-90	1	1	0	0	0	[32]
91		C-91	0	1	1	1	0	[33]
92		C-92	1	1	0	0	1	[34]
93	S1PR5	C-93	0	1	0	0	1	[34]
94		C-94	1	0	0	0	0	[32]
95		C-95	1	0	0	1	0	[32]
96		C-96	1	0	0	1	0	[32]
97		C-97	0	0	0	1	0	[32]
98		C-98	1	0	0	0	0	[32]
99		C-99	1	0	1	1	0	[32]
100		C-100	1	0	0	1	1	[32]

Table S2. The SMILES and Compound ID of compounds in Table S1.

SMILES	Compound ID
CC\C(=C(/c1ccccc1)c1ccc(OCCN(C)C)cc1)c1ccccc1	C-01
OC(=O)c1ccc(CCN2CCN(C\C=C\c3ccc(Cl)cc3)CC2)cc1	C-02
COc1ccc(CCN2CCc3cc(O)c(OC)cc3C2)cc1OC	C-03
COc1cccc(CCN2CCc3cc(O)c(OC)cc3C2)c1	C-04
COc1cc2CN(CCc3ccc(Cl)cc3)CCc2cc1O	C-05
COc1cccc2C(=O)c3c(O)c4C[C@](O)(C[C@H](O[C@H]5C[C@H](N)[C@H](O)[C@H](C)O5)c4c(O)c3C(=O)c12)C(=O)CO	C-06
COc1cccc2C(=O)c3c(O)c4C[C@](O)(C[C@H](O[C@H]5C[C@H](NCNC(=O)c6ccccc6O)[C@H](O)[C@H](C)O5)c4c(O)c3C(=O)c12)C(=O)CO	C-07
CC\C(=C(\c1ccc(O)cc1)c1ccc(OCCN(C)CCOCCOCCO\N=C\c2ccc(O)c(c2)C(=O)NCN[C@H]2C[C@H](O[C@H]3C[C@@](O)(Cc4c(O)c5C(=O)c6cccc(OC)c6C(=O)c5c(O)c34)C(=O)CO)O[C@@H](C)[C@H]2O)cc1)c1ccccc1	C-08
CN(C)C1CCc2[nH]c3ccc(NC(=O)c4ccc(F)cc4)cc3c2C1	C-09
Cc1ccc2c(OCCN3CCC(Cc4cccc(c4)N4CCCS4(=O)=O)CC3)cccc2n1	C-10
Cc1ccc2c(OCCN3CCC(Cc4cccc(NS(C)(=O)=O)c4)CC3)cccc2n1	C-11
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)c1Cl	C-12
CN1CCC(CC1)C(=O)c1ccc(F)c(NC(=O)c2ccc(F)cc2)c1	C-13
CN1CCC(CC1)C(=O)c1cc(F)cc(NC(=O)c2ccc(F)cc2)c1	C-14
CN1CCC(CC1)C(=O)c1cc(NC(=O)c2ccc(F)cc2)ccc1F	C-15
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)c1F	C-16
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)c1	C-17
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)c1O	C-18
COc1c(NC(=O)c2ccc(F)cc2)cccc1C(=O)C1CCN(C)CC1	C-19
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)c1C	C-20
CN1CCC(CC1)c1c[nH]c2ccc(NC(=O)c3ccc(F)cc3)cc12	C-21
CN1CCC(CC1)C(=O)c1cccc(NC(=O)c2ccc(F)cc2)n1	C-22
CCCCC/C=C\C[C@H](/C=C/C=C\C=C\[C@H](CCCC(=O)O)O)O	C-23
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccncc2)-c2ccncc2)c1CCC(O)=O	C-24
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(n2)-c2ccccc2)-c2ccccc2)c1CCC(O)=O	C-25
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccc3OCCOc3c2)-c2ccc(F)cc2)c1CCC(O)=O	C-26
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccc3OCOc3c2)-c2ccsc2)c1CCC(O)=O	C-27
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2cccc(F)c2)-c2ccc3OCOc3c2)c1CCC(O)=O	C-28
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccc3OCCOc3c2)-c2ccc3OCCOc3c2)c1CCC(O)=O	C-29
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2cccc(F)c2)-c2ccncc2)c1CCC(O)=O	C-30
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccncc2)-c2ccsc2)c1CCC(O)=O	C-31
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccccc2)-c2ccccc2)c1CCC(O)=O	C-32
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccc3OCOc3c2)-c2ccncc2)c1CCC(O)=O	C-33

OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccnc(Cl)c2)-c2ccsc2)c1CCC(O)=O	C-34
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccc3OCOc3c2)-c2cncnc2)c1CCC(O)=O	C-35
OC(=O)CCCOc1cccc(CCCCCCOc2cc(cc(c2)-c2ccsc2)-c2ccsc2)c1CCC(O)=O	C-36
OC(=O)CCCn1cc(CC(O)=O)c2c(\C=C\c3ccc(OCCCCOc4ccccc4)cc3)cccc12	C-37
OC(=O)CCCN1CC(Oc2c(NC(=O)c3ccc(OCCCCc4ccccc4)cc3)cccc12)C(O)=O	C-38
OC(=O)C1CCCC(C1)NC(=O)c1cc(ccc1OCCCc1ccc(OCCCCOC2CCCCC2)cc1)C(O)=O	C-39
O=C(Nc1cccc2c1oc(cc2=O)-c1nnn[nH]1)c1ccc(OCCCCc2ccccc2)cc1	C-40
COc1cc(ccc1Cc1cn(C)c2ccc(NC(=O)OC3CCCC3)cc12)C(=O)NS(=O)(=O)c1ccccc1C	C-41
OC(=O)CCCN1C[C@@H](Oc2c(NC(=O)c3ccc(OCCCCc4ccccc4)cc3)cccc12)C(O)=O	C-42
OC(=O)CCCN1C[C@H](Oc2c(NC(=O)c3ccc(OCCCCc4ccccc4)cc3)cccc12)C(O)=O	C-43
CCN(CC)C(=O)\C=C(/C)c1ccc2oc(C(=O)c3ccc(cc3)C#N)c(NC(=O)C(\C#N)=C(\C)O)c2c1	C-44
CCCCC\C=C/C\C=C/C=C/C=C/[C@@H](Sc1ccc(cc1)C(O)=O)[C@@H](O)CCCC(O)=O	C-45
OC(=O)CCCc1cn(CC(O)=O)c2c(\C=C\c3ccc(OCCCCOc4ccccc4)cc3)cccc12	C-46
OC(=O)CCCc1cn(CC(O)=O)c2c(\C=C\c3ccc(OCCCCc4ccccc4)cc3)cccc12	C-47
OC(=O)CCCc1cn(CC(O)=O)c2c(\C=C\c3ccc(OCCCCc4cccc(Cl)c4)cc3)cccc12	C-48
OC(=O)CCCc1cn(CC(O)=O)c2c(\C=C\c3ccc(OCCCCc4c(F)ccc(F)c4F)cc3)cccc12	C-49
Cc1c(F)cccc1CCCCOc1ccc(\C=C\c2cccc3c(CCCC(O)=O)cn(CC(O)=O)c23)cc1	C-50
Cc1c(Cl)cccc1CCCCOc1ccc(\C=C\c2cccc3c(CCCC(O)=O)cn(CC(O)=O)c23)cc1	C-51
N[C@H](C(=O)O)Cn1cc(C)c(=O)n(c1=O)Cc1ccsc1C(=O)O	C-52
N[C@H](CCC(O)=O)C(O)=O	C-53
CC(=C)[C@H]1CN[C@@H]([C@H]1CC(O)=O)C(O)=O	C-54
COc1ccccc1\C=C\C[C@H](C[C@H](N)C(O)=O)C(O)=O	C-55
N[C@@H](C[C@@H](C\C=C\c1ccc(Cl)cc1)C(O)=O)C(O)=O	C-56
N[C@@H](C[C@@H](C\C=C\c1ccc2ccccc2c1)C(O)=O)C(O)=O	C-57
N[C@@H](Cn1ccc(=O)n(Cc2ccccc2C(O)=O)c1=O)C(O)=O	C-58
OC(=O)[C@@H]1CC(F)(F)CN1C[C@H]1CC[C@H]2CN[C@@H](C[C@H]2C1)C(O)=O	C-59
C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CC[C@@H]2O	C-60
CC(=O)c1ccc2N[C@H]([C@H]3CC=C[C@H]3c2c1)c1cc2OCOc2cc1Br	C-61
C[C@@H]1Cc2cc(ccc2[C@H](N1CC(C)(C)F)c1c(F)cc(\C=C\C(O)=O)cc1F)C#Cc1cnn(C)c1	C-62
C[C@]12CCC3C(CCc4cc(O)ccc34)C1CC[C@@]2(O)C#Cc1ccc(cc1)[N+](C)(C)C	C-63
Brc1cc2OCOc2cc1[C@@H]1Nc2ccccc2[C@@H]2C=CC[C@H]12	C-64
C[C@]12CCC3C(CCc4cc(O)ccc34)C1CC[C@@]2(O)C#Cc1cccc(c1)C([O-])=O	C-65
C[C@]12CCC3C(CCc4cc(O)ccc34)C1CC[C@@]2(O)C#Cc1ccc(C[NH3+])cc1	C-66
CC(C)(C)OC(=O)NCc1ccc(cc1)C#C[C@]1(O)CCC2C3CCc4cc(O)ccc4C3CC[C@]12C	C-67
Brc1cc2OCOc2cc1[C@@H]1Nc2ccccc2[C@@H]2C=CC[C@H]12	C-68
C[C@]12CC[C@H]3[C@@H](CCc4cc(O)ccc34)[C@@H]1CC[C@@H]2O	C-69
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccc(Cl)c(F)c2)CC1	C-70
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccc(F)c(Cl)c2)CC1	C-71
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccc(F)c(F)c2)CC1	C-72
COc1ccc(cc1)N(C(=O)OC[C@H]1CC[C@H](COCC(O)=O)CC1)c1cccc(F)c1	C-73
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccc(F)cc2)c2cccc(F)c2)CC1	C-74

CC(C)N(CCCCOCC(O)=O)c1cnc(-c2ccccc2)c(n1)-c1ccccc1	C-75
OC(=O)COC[C@H]1CC[C@@H](COC(=O)N(c2ccccc2)c2ccccc2)CC1	C-76
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccccc2)CC1	C-77
CC(C)N(CCCCOCC(=O)NS(C)(=O)=O)c1cnc(-c2ccccc2)c(n1)-c1ccccc1	C-78
OC(=O)COC[C@H]1CC[C@@H](COC(=O)N(c2ccccc2)c2ccc(Cl)cc2)CC1	C-79
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccc(Cl)cc2)CC1	C-80
OC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2cccc(F)c2)CC1	C-81
CC#CCC(C)[C@H](O)\C=C\[C@H]1[C@H](O)C[C@@H]2C\C(C[C@H]12)=C/CCCC(O)=O	C-82
OS(=O)(=O)CCNC(=O)COC[C@H]1CC[C@H](COC(=O)N(c2ccccc2)c2ccc(Cl)cc2)CC1	C-83
COc1ccc(cc1)N(C(=O)OC[C@H]1CC[C@H](COCC(O)=O)CC1)c1ccccc1	C-84
OC(=O)COC[C@H]1CC[C@@H](COC(=O)N(c2ccccc2)c2cccc(F)c2)CC1	C-85
CCCCCCCCc1cccc(c1)C1CC(C1)(N)COP(=O)(O)O	C-86
CCC\N=C1/S\C(=C/c2ccc(OC[C@@H](O)CO)c(Cl)c2)C(=O)N1c1ccccc1C	C-87
CN(C)CC[C@H](N(C)C(=O)c1c(C)cc(cc1C)-c1cccc(NS(=O)(=O)c2cc(C)c(Cl)cc2C)c1)C(O)=O	C-88
OC(=O)Cc1cccc(NCc2cc3cc(ccc3s2)-c2ccc3ccccc3c2)c1	C-89
CCc1ccccc1-c1cccc2sc(CNCc3ccc(cc3)C(O)=O)cc12	C-90
CCCCCCCCc1ccc(CCC(N)(CO)COP(O)(O)=O)cc1	C-91
CC(C)Cc1cc(cc(C)n1)-c1nc(no1)-c1cc(C)c(OC[C@@H]2CCC(=O)N2)c(C)n1	C-92
CC(C)Cc1cc(nc(C)n1)-c1nc(no1)-c1ccc(O[C@@H]2CCC(=O)NC2)c(F)c1	C-93
CCc1ccccc1-c1cccc2sc(CNc3ccc(CC(O)=O)cc3)cc12	C-94
NC(CNCc1cc2cc(ccc2s1)-c1ccc2ccccc2c1)C(O)=O	C-95
NC(CNCc1cc2cccc(-c3ccc4ccccc4c3)c2s1)C(O)=O	C-96
CC(CNCc1cc2cccc(-c3ccc4ccccc4c3)c2s1)C(O)=O	C-97
CCc1ccccc1-c1cccc2cc(CNc3cccc(CC(O)=O)c3)sc12	C-98
OC(=O)CCCNCc1cc2cccc(-c3cccc4ccccc34)c2s1	C-99
CCc1ccccc1-c1cccc2cc(CNCCC(O)=O)sc12	C-100

Table S3. The optimal hyper-parameters of 30 different models.

Model		The optimal hyper-parameters
Des	Bagging	n_estimators: 218
	DT	criterion: gini
	GBDT	learning_rate': 0.5; n_estimators: 1960
	KNN	n_neighbors: 7, weights: uniform
	SVM	C: 80
FP4	Bagging	n_estimators: 228
	DT	criterion: gini
	GBDT	learning_rate: 0.4, n_estimators: 1925
	KNN	n_neighbors: 5, weights: uniform
	SVM	C: 500
KR	Bagging	n_estimators: 228
	DT	criterion: entropy
	GBDT	learning_rate: 0.2, n_estimators: 1760
	KNN	n_neighbors: 3, weights: uniform
	SVM	C: 180
MACCS	Bagging	n_estimators: 235
	DT	criterion: gini
	GBDT	learning_rate: 0.4, n_estimators: 1820
	KNN	n_neighbors: 5, knn__weights: distance
	SVM	C: 100
Morgan	Bagging	n_estimators: 215
	DT	criterion: entropy
	GBDT	learning_rate: 0.5, n_estimators: 1775
	KNN	n_neighbors: 5, weights: uniform
	SVM	C: 80
PubChem	Bagging	n_estimators: 246
	DT	criterion: gini
	GBDT	learning_rate: 0.4, n_estimators: 1735
	KNN	n_neighbors: 5, weights: distance
	SVM	C: 100

Table S4. Ten-fold cross validation results of different models. All values were in percentage.

Model		F1	ACC	NPV	PPV (P)	SP	SE (R)
Des	Bagging	82.95±0.99	86.01±0.59	86.25±0.58	85.65±1.22	90.11±0.74	80.42±1.21
	DT	77.29±1.10	80.68±0.78	83.47±0.25	76.94±1.71	82.91±1.16	77.65±0.65
	GBDT	**84.00±0.77**	**86.69±0.56**	**87.49±0.65**	**85.52±1.27**	**89.74±0.79**	**82.54±0.81**
	KNN	78.10±0.68	80.16±0.50	86.54±0.61	73.32±1.13	77.66±0.80	83.57±0.60
	SVM	83.75±0.63	86.35±0.55	87.69±0.92	84.47±0.77	88.79±0.45	83.05±1.01
FP4	Bagging	80.95±0.58	84.55±0.31	84.46±0.44	84.68±0.94	89.69±0.58	77.55±0.67
	DT	76.23±1.01	79.71±0.75	82.77±0.50	75.65±1.45	81.83±0.96	76.82±0.65
	GBDT	**81.99±0.62**	**85.06±0.40**	**85.97±0.53**	**83.72±1.23**	**88.53±0.75**	**80.35±0.57**
	KNN	80.13±0.87	82.19±0.61	87.82±0.45	75.93±1.39	80.24±1.05	84.85±0.77
	SVM	80.41±0.93	83.50±0.63	85.41±0.56	80.84±1.31	86.08±0.75	79.98±0.80
KR	Bagging	82.75±0.53	85.94±0.33	85.84±0.55	86.09±0.94	90.54±0.63	79.67±0.81
	DT	77.43±1.23	80.84±0.73	83.54±0.53	77.19±1.42	83.15±0.78	77.68±1.30
	GBDT	83.71±0.65	86.45±0.41	87.28±0.43	85.25±1.17	89.55±0.75	82.24±0.65
	KNN	80.76±0.76	82.70±0.54	88.51±0.46	76.30±0.98	80.42±0.90	85.78±0.85
	SVM	**84.96±0.58**	**87.33±0.34**	**88.72±0.43**	**85.40±1.03**	**89.38±0.68**	**84.53±0.65**
MACCS	Bagging	82.88±0.79	86.00±0.61	86.05±0.57	85.92±1.30	90.36±0.90	80.07±0.75
	DT	76.66±0.94	80.03±0.75	83.18±0.57	75.90±1.70	81.92±1.21	77.46±0.51
	GBDT	83.15±0.64	86.02±0.42	86.79±0.50	84.88±1.16	89.34±0.71	81.50±0.61
	KNN	80.93±0.50	82.93±0.41	88.38±0.44	76.83±1.10	81.05±0.94	85.50±0.44
	SVM	**83.27±0.74**	**85.95±0.49**	**87.39±0.61**	**83.92±1.11**	**88.38±0.66**	**82.64±0.84**
Morgan	Bagging	83.16±0.83	86.17±0.55	86.39±0.55	85.83±1.09	90.23±0.66	80.65±0.88
	DT	77.75±0.86	81.07±0.60	83.82±0.78	77.39±1.04	83.23±0.73	78.13±1.13
	GBDT	83.94±0.64	86.56±0.52	87.68±0.69	84.96±0.90	89.22±0.57	82.95±0.77
	KNN	80.67±0.75	82.48±0.51	88.81±0.46	75.71±1.24	79.65±0.94	86.33±0.72
	SVM	**85.55±0.46**	**87.86±0.28**	**89.02±0.51**	**86.24±0.81**	**90.05±0.49**	**84.89±0.62**
PubChem	Bagging	82.97±0.88	86.07±0.61	86.12±0.52	85.99±1.45	90.40±0.97	80.16±0.90
	DT	77.73±0.87	80.99±0.69	83.89±0.56	77.16±1.44	82.97±1.11	78.31±0.70
	GBDT	84.05±0.62	86.74±0.43	87.47±0.55	85.69±0.99	89.88±0.65	82.48±0.74
	KNN	80.87±0.56	82.88±0.36	88.34±0.42	76.75±1.06	80.98±0.84	85.46±0.59
	SVM	**84.64±0.62**	**87.10±0.35**	**88.36±0.46**	**85.32±0.73**	**89.39±0.45**	**83.97±0.81**

$$ACC = \frac{TP + TN}{TP + TN + FP + FN};\quad NPV = \frac{TN}{TN + FN};\quad SP = \frac{TN}{TN + FP};$$

ACC: Accuracy; NPV: Negative Predictive Value; SP: Specificity; PPV: Positive Predictive Value (= Precision); SE: Sensitivity (= Recall).

Table S5. Test set validation results of different models. All values were in percentage.

Model		F1	ACC	NPV	PPV(P)	SP	SE(R)
Des	Bagging	82.01	85.24	85.62	84.68	89.45	79.51
	DT	76.06	79.65	82.59	75.70	82.01	76.42
	GBDT	83.31	85.99	87.43	83.97	88.43	82.66
	KNN	78.35	80.37	86.86	73.44	77.73	83.96
	SVM	**83.82**	**86.30**	**88.16**	**83.76**	**88.07**	**83.88**
FP4	Bagging	80.07	83.85	83.90	83.78	89.11	76.68
	DT	75.75	79.39	82.36	75.41	81.80	76.11
	GBDT	**82.08**	**85.15**	**86.05**	**83.83**	**88.62**	**80.41**
	KNN	80.52	82.57	88.12	76.35	80.65	85.18
	SVM	80.00	83.20	85.06	80.61	85.99	79.40
KR	Bagging	82.50	85.79	85.56	86.16	90.68	79.14
	DT	76.78	80.13	83.33	75.95	81.97	77.64
	KNN	80.19	82.25	87.87	75.98	80.31	84.89
	GBDT	83.37	86.04	87.46	84.05	88.49	82.69
	SVM	**84.37**	**86.86**	**88.25**	**84.92**	**89.09**	**83.82**
MACCS	Bagging	82.58	85.76	85.88	85.56	90.13	79.80
	DT	76.99	80.51	83.15	76.93	83.05	77.06
	GBDT	82.30	85.23	86.46	83.48	88.22	81.16
	KNN	80.79	82.85	88.24	76.77	81.08	85.26
	SVM	**83.27**	**85.93**	**87.46**	**83.79**	**88.26**	**82.75**
Morgan	Bagging	81.99	85.15	85.80	84.18	88.98	79.92
	DT	77.72	80.84	84.20	76.51	82.22	78.97
	GBDT	83.72	86.33	87.72	84.39	88.73	83.07
	KNN	80.57	82.40	88.77	75.58	79.55	86.28
	SVM	**85.11**	**87.52**	**88.67**	**85.90**	**89.85**	**84.34**
PubChem	Bagging	82.42	85.61	85.82	85.29	89.91	79.75
	DT	76.24	79.73	82.82	75.63	81.84	76.86
	GBDT	83.27	85.99	87.27	84.17	88.64	82.38
	KNN	80.73	82.80	88.16	76.74	81.08	85.15
	SVM	**84.77**	**87.16**	**88.66**	**85.09**	**89.15**	**84.46**

References

1. J. R. Quinlan, *Machine Learning*, 1986, **1**, 81-106.
2. L. Breiman, *Machine Learning*, 1996, **24**, 123-140.
3. J. H. Friedman, *Annals of Statistics*, 2001, **29**, 1189-1232.
4. T. M. Cover and P. E. Hart, *IEEE Transactions on Information Theory*, 1967, **13**, 21-27.
5. W. M. Czarnecki, S. Podlewska and A. J. Bojarski, *J. Cheminf.*, 2015, **7**, 38.
6. C. J. C. Burges, *Data Mining and Knowledge Discovery*, 1998, **2**, 121-167.
7. V. C. Jordan, *Journal of medicinal chemistry*, 2003, **46**, 883-908.
8. A. Horling, C. Müller, R. Barthel, F. Bracher and P. Imming, *Journal of Medicinal Chemistry*, 2012, **55**, 7614-7622.
9. P. J. Burke, B. T. Kalet and T. H. Koch, *Journal of Medicinal Chemistry*, 2004, **47**, 6509-6518.
10. Bret D. Wallace, Adam B. Roberts, Rebecca M. Pollet, James D. Ingle, Kristen A. Biernat, Samuel J. Pellock, Madhu K. Venkatesh, L. Guthrie, Sara K. O'Neal, Sara J. Robinson, M. Dollinger, E. Figueroa, Sarah R. McShane, Rachel D. Cohen, J. Jin, Stephen V. Frye, William C. Zamboni, C. Pepe-Ranney, S. Mani, L. Kelly and Matthew R. Redinbo, *Chemistry & Biology*, 2015, **22**, 1238-1249.
11. S. E. Ward, P. J. Eddershaw, C. M. Scott, L. J. Gordon, P. J. Lovell, S. H. Moore, P. W. Smith, K. R. Starr, K. M. Thewlis and J. M. Watson, *Journal of Medicinal Chemistry*, 2008, **51**, 2887-2890.
12. D. Zhang, M.-J. Blanco, B.-P. Ying, D. Kohlman, S. X. Liang, F. Victor, Q. Chen, J. Krushinski, S. A. Filla, K. J. Hudziak, B. M. Mathes, M. P. Cohen, D. Zacherl, D. L. G. Nelson, D. B. Wainscott, S. E. Nutter, W. H. Gough, J. M. Schaus and Y.-C. Xu, *Bioorg. Med. Chem. Lett.*, 2015, **25**, 4337-4341.
13. S.-K. Choi, D. Green, A. Ho, U. Klein, D. Marquess, R. Taylor and S. D. Turner, *Journal of Medicinal Chemistry*, 2008, **51**, 3609-3616.
14. T. Yokomizo, T. Izumi, K. Chang, Y. Takuwa and T. Shimizu, *Nature*, 1997, **387**, 620-624.
15. R. A. Goodnow, A. Hicks, A. Sidduri, A. Kowalczyk, R. Dominique, Q. Qiao, J. P. Lou, P. Gillespie, N. Fotouhi, J. Tilley, N. Cohen, S. Choudhry, G. Cavallo, S. A. Tannu, J. D. Ventre, D. Lavelle, N. S. Tare, H. Oh, M. Lamb, G. Kurylko, R. Hamid, M. B. Wright, A. Pamidimukkala, T. Egan, U. Gubler, A. F. Hoffman, X. Wei, Y. L. Li, J. O'Neil, R. Marcano, K. Pozzani, T. Molinaro, J. Santiago, L. Singer, M. Hargaden, D. Moore, A. R. Catala, L. C. F. Chao, G. Hermann, R. Venkat, H. Mancebo and L. M. Renzetti, *Journal of Medicinal Chemistry*, 2010, **53**, 3502-3516.
16. S. Itadani, K. Yashiro, Y. Aratani, T. Sekiguchi, A. Kinoshita, H. Moriguchi, N. Ohta, S. Takahashi, A. Ishida, Y. Tajima, K. Hisaichi, M. Ima, J. Ueda, H. Egashira, T. Sekioka, M. Kadode, Y. Yonetomi, T. Nakao, A. Inoue, H. Nomura, T. Kitamine, M. Fujita, T. Nabe, Y. Yamaura, N. Matsumura, A. Imagawa, Y. Nakayama, J. Takeuchi and K. Ohmoto, *Journal of Medicinal Chemistry*, 2015, **58**, 6093-6113.
17. S. Itadani, S. Takahashi, M. Ima, T. Sekiguchi, M. Fujita, Y. Nakayama and J. Takeuchi, *ACS Medicinal Chemistry Letters*, 2014, **5**, 1230-1234.
18. P. T. Atlason, C. L. Scholefield, R. J. Eaves, M. B. Mayo-Martin, D. E. Jane and E. Molnár, *Mol Pharmacol*, 2010, **78**, 1036-1045.
19. C. Pedregal, I. Collado, A. Escribano, J. Ezquerra, C. Domínguez, A. I. Mateo, A. Rubio, S. R. Baker, J. Goldsworthy, R. K. Kamboj, B. A. Ballyk, K. Hoo and D. Bleakman, *Journal of Medicinal Chemistry*, 2000, **43**, 1958-1968.
20. E. Szymańska, P. Chałupnik, K. Szczepańska, A. M. Cuñado Moral, D. S. Pickering, B. Nielsen, T.

N. Johansen and K. Kieć-Kononowicz, *Bioorg. Med. Chem. Lett.*, 2016, **26**, 5568-5572.

21. N. Krogsgaard-Larsen, C. G. Delgar, K. Koch, P. M. G. E. Brown, C. Møller, L. Han, T. H. V. Huynh, S. W. Hansen, B. Nielsen, D. Bowie, D. S. Pickering, J. S. Kastrup, K. Frydenvang and L. Bunch, *Journal of Medicinal Chemistry*, 2017, **60**, 441-457.
22. C. G. Bologa, C. M. Revankar, S. M. Young, B. S. Edwards, J. B. Arterburn, A. S. Kiselyov, M. A. Parker, S. E. Tkachenko, N. P. Savchuck, L. A. Sklar, T. I. Oprea and E. R. Prossnitz, *Nat. Chem. Biol.*, 2006, **2**, 207-212.
23. D. M. Huryn, L. O. Resnick and P. Wipf, *Journal of Medicinal Chemistry*, 2013, **56**, 7161-7176.
24. C. Ramesh, T. K. Nayak, R. Burai, M. K. Dennis, H. J. Hathaway, L. A. Sklar, E. R. Prossnitz and J. B. Arterburn, *Journal of Medicinal Chemistry*, 2010, **53**, 1004-1014.
25. G. Scapin, S. B. Patel, C. Chung, J. P. Varnerin, S. D. Edmondson, A. Mastracchio, E. R. Parmee, S. B. Singh, J. W. Becker, L. H. T. Van der Ploeg and M. R. Tota, *Biochemistry*, 2004, **43**, 6091-6100.
26. C. M. Revankar, H. D. Mitchell, A. S. Field, R. Burai, C. Corona, C. Ramesh, L. A. Sklar, J. B. Arterburn and E. R. Prossnitz, *ACS Chem. Biol.*, 2007, **2**, 536-544.
27. M. K. Dennis, R. Burai, C. Ramesh, W. K. Petrie, S. N. Alcon, T. K. Nayak, C. G. Bologa, A. Leitao, E. Brailoiu, E. Deliu, N. J. Dun, L. A. Sklar, H. J. Hathaway, J. B. Arterburn, T. I. Oprea and E. R. Prossnitz, *Nat. Chem. Biol.*, 2009, **5**, 421-427.
28. T.-A. Tran, B. Kramer, Y.-J. Shin, P. Vallar, P. D. Boatman, N. Zou, C. R. Sage, T. Gharbaoui, A. Krishnan, B. Pal, S. R. Shakya, A. Garrido Montalban, J. W. Adams, J. Ramirez, D. P. Behan, A. Shifrina, A. Blackburn, T. Leakakos, Y. Shi, M. Morgan, A. Sadeque, W. Chen, D. J. Unett, I. Gaidarov, X. Chen, S. Chang, H.-H. Shu, S.-F. Tung and G. Semple, *Journal of Medicinal Chemistry*, 2017, **60**, 913-927.
29. P. C. Kennedy, R. Zhu, T. Huang, J. L. Tomsig, T. P. Mathews, M. David, O. Peyruchaud, T. L. Macdonald and K. R. Lynch, *J Pharmacol Exp Ther*, 2011, **338**, 879-889.
30. M. H. Bolli, S. Abele, C. Binkert, R. Bravo, S. Buchmann, D. Bur, J. Gatfield, P. Hess, C. Kohl, C. Mangold, B. Mathys, K. Menyhart, C. Müller, O. Nayler, M. Scherz, G. Schmidt, V. Sippel, B. Steiner, D. Strasser, A. Treiber and T. Weller, *Journal of Medicinal Chemistry*, 2010, **53**, 4198-4211.
31. D. Angst, P. Janser, J. Quancard, P. Buehlmayer, F. Berst, L. Oberer, C. Beerli, M. Streiff, C. Pally, R. Hersperger, C. Bruns, F. Bassilana and B. Bollbuck, *Journal of Medicinal Chemistry*, 2012, **55**, 9722-9734.
32. W. Hur, H. Rosen and N. S. Gray, *Bioorg. Med. Chem. Lett.*, 2017, **27**, 1-5.
33. R. Albert, K. Hinterding, V. Brinkmann, D. Guerini, C. Müller-Hartwieg, H. Knecht, C. Simeon, M. Streiff, T. Wagner, K. Welzenbach, F. Zécri, M. Zollinger, N. Cooke and E. Francotte, *Journal of Medicinal Chemistry*, 2005, **48**, 5373-5377.
34. J. C. Horan, D. Kuzmich, P. Liu, D. DiSalvo, J. Lord, C. Mao, T. D. Hopkins, H. Yu, C. Harcken, R. Betageri, M. Hill-Drzewi, L. Patenaude, M. Patel, K. Fletcher, D. Terenzzio, B. Linehan, H. Xia, M. Patel, D. Studwell, C. Miller, E. Hickey, J. I. Levin, D. Smith, R. A. Kemper, L. K. Modis, L. C. Bannen, D. S. Chan, M. B. Mac, S. Ng, Y. Wang, W. Xu and R. M. Lemieux, *Bioorg. Med. Chem. Lett.*, 2016, **26**, 466-471.