Supporting Information

Machine learning-based models for high-throughput classification of human pregnane X receptor activators

Yiyuan Gou†, ‡, Lilai Shen†, Shixuan Cui†, ‡, *, Meiling Huang†, Yiqu Wu†, Penghan Li†, Shulin Zhuang†, ‡, *

[†] Key Laboratory of Environment Remediation and Ecological Health, Ministry of Education, College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China.

[‡] Women's Reproductive Health Key Laboratory of Zhejiang Province, Women's Hospital, School of Medicine, Zhejiang University, Hangzhou 310006, China.

* Corresponding author: College of Environmental and Resource Sciences, Zhejiang University, Hangzhou 310058, China. Email address: sxcui@zju.edu.cn (S Cui) or shulin@zju.edu.cn (S Zhuang).

This file contains 1 text, 6 tables and 1 figure.

Text S1. Model evaluation metrics

Balanced accuracy reflects the accuracy of the model especially when data sets are imbalanced. Precision measures the ability of the classifier not to label as positive a sample that is negative, and recall measures the ability of the classifier to find all the positive samples. The F1 score is the harmonic mean of the precision and recall. Receiver operating characteristic curve (ROC) can graphically present the model performance in a visual way. The area under the curve (AUC) shows the ability of the model to separate classes. Another metric of model classification performance is MCC, which only generates a high score when both positive and negative instances are correctly predicted. Cohen's Kappa (CK) is used to measure overall model performance by normalizing the accuracy to the probability that the classification would agree.

$$\text{balanced accuracy} = \frac{\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} + \frac{N_{\mathrm{TN}}}{N_{\mathrm{TN}} + N_{\mathrm{FP}}}}{2}$$

$$\text{precision} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}}$$

$$\text{recall} = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}}$$

$$\text{F1 score} = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

$$\mathrm{CK} = \frac{\text{Accuracy} - p_e}{1 - p_e}$$

$$\mathrm{MCC} = \frac{N_{\mathrm{TP}} \cdot N_{\mathrm{TN}} - N_{\mathrm{FP}} \cdot N_{\mathrm{FN}}}{\sqrt{(N_{\mathrm{TP}} + N_{\mathrm{FP}})(N_{\mathrm{TP}} + N_{\mathrm{FN}})(N_{\mathrm{TN}} + N_{\mathrm{FP}})(N_{\mathrm{TN}} + N_{\mathrm{FN}})}}$$

where N_{TP}, N_{TN}, N_{FP}, and N_{FN} represent the number of true positive, true negative, false positive, and false negative, respectively. $p_e = p_{\mathrm{True}} + p_{False}$, where

$p_{\mathrm{True}} = \frac{N_{\mathrm{TP}}+N_{\mathrm{FN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}} \cdot \frac{N_{\mathrm{TP}}+N_{\mathrm{FP}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}$, $p_{\mathrm{False}} = \frac{N_{\mathrm{TN}}+N_{\mathrm{FN}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}} \cdot \frac{N_{\mathrm{TN}}+N_{\mathrm{FP}}}{N_{\mathrm{TP}}+N_{\mathrm{TN}}+N_{\mathrm{FP}}+N_{\mathrm{FN}}}$

Table S1 The brief definitions of 87 selected descriptors

Descriptor	Type	Description
PEOE_VSA7	PEOE_VSA	MOE Charge VSA Descriptor 7 (-0.05 <= x < 0.00)
BCUT2D_CHGHI	BCUT descriptor	highest eigenvalue weighted by gasteiger charges
VSA_EState3	VSA_Estate	VSA EState Descriptor 3 (5.00 <= x < 5.41)
VSA_EState2	VSA_Estate	VSA EState Descriptor 2 (4.78 <= x < 5.00)
EState_VSA3	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
VSA_EState10	VSA_Estate	VSA EState Descriptor 10 (11.00 <= x < inf)
FpDensityMorgan1	FpDensityMorgan	Morgan fingerprint density
Chi4n	Chi indices	Similar to Hall Kier Chi4v, but uses nVal instead of valence.This makes a big difference after we get out of the first row.Rev. Comput. Chem. 2:367-422 (1991).
SlogP_VSA5	SlogP_VSA	MOE logP VSA Descriptor 5 (0.10 <= x < 0.15)
BCUT2D_LOGPHI	BCUT descriptor	highest eigenvalue weighted by crippen logP
BCUT2D_CHGLO	BCUT descriptor	lowest eigenvalue weighted by gasteiger charges
BCUT2D_MWLOW	BCUT descriptor	lowest eigenvalue weighted by atomic masses
BCUT2D_LOGPLOW	BCUT descriptor	lowest eigenvalue weighted by crippen logP
MinAbsPartialCharge	Partial Charge	Returns molecular charge descriptors
BCUT2D_MWHI	BCUT descriptor	highest eigenvalue weighted by atomic masses
MaxAbsPartialCharge	Partial Charge	Returns molecular charge descriptors
PEOE_VSA11	PEOE_VSA	MOE Charge VSA Descriptor 11 (0.15 <= x < 0.20)
BCUT2D_MRHI	BCUT descriptor	highest eigenvalue weighted by crippen MRR
PEOE_VSA9	PEOE_VSA	MOE Charge VSA Descriptor 9 (0.05 <= x < 0.10)
BalabanJ	Balaban's J index	Balaban's J value for a molecule,Chem. Phys. Lett. 89:399-404 (1982).
SMR_VSA5	SMR_VSA	MOE MR VSA Descriptor 5 (2.45 <= x < 2.75)
Chi4v	Chi indices	From equations (5),(15) and (16) of Rev. Comp. Chem. vol 2, 367-422, (1991)
MolMR	MolMR	Wildman-Crippen MR value.Wildman and Crippen JCICS 39:868-73 (1999)
PEOE_VSA14	PEOE_VSA	MOE Charge VSA Descriptor 14 (0.30 <= x < inf)
SMR_VSA9	SMR_VSA	MOE MR VSA Descriptor 9 (3.80 <= x < 4.00)
MinAbsEStateIndex	Estate Index	Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS _31_ 76-81 (1994)
EState_VSA1	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
BCUT2D_MRLOW	BCUT descriptor	lowest eigenvalue weighted by crippen MRR
Kappa2	Kappa descriptors	Hall-Kier Kappa2 value
SMR_VSA3	SMR_VSA	MOE MR VSA Descriptor 3 (1.82 <= x < 2.24)
VSA_EState9	VSA_Estate	VSA EState Descriptor 9 (7.00 <= x < 11.00)
EState_VSA6	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
FpDensityMorgan2	FpDensityMorgan	Morgan fingerprint density
PEOE_VSA8	PEOE_VSA	MOE Charge VSA Descriptor 8 (0.00 <= x < 0.05)
EState_VSA8	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
EState_VSA4	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
FpDensityMorgan3	FpDensityMorgan	Morgan fingerprint density
SlogP_VSA1	SlogP_VSA	MOE logP VSA Descriptor 1 (-inf < x < -0.40)
SMR_VSA1	SMR_VSA	MOE MR VSA Descriptor 1 (-inf < x < 1.29)
PEOE_VSA2	PEOE_VSA	MOE Charge VSA Descriptor 2 (-0.30 <= x < -0.25)

Table S1(continued) The brief definitions of 87 selected descriptors

Descriptor	Type	Description
MaxAbsEStateIndex	Estate Index	Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS _31_ 76-81 (1993)
HallKierAlpha	HallKierAlpha	The Hall-Kier alpha value for a molecule.Rev. Comput. Chem. 2:367-422 (1991).
VSA_EState4	VSA_Estate	VSA EState Descriptor 4 (5.41 <= x < 5.74)
EState_VSA2	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
SlogP_VSA3	SlogP_VSA	MOE logP VSA Descriptor 3 (-0.20 <= x < 0.00)
fr_benzene	fr_benzene	Number of benzene rings
VSA_EState8	VSA_Estate	VSA EState Descriptor 8 (6.45 <= x < 7.00)
SlogP_VSA2	SlogP_VSA	MOE logP VSA Descriptor 2 (-0.40 <= x < -0.20)
PEOE_VSA6	PEOE_VSA	MOE Charge VSA Descriptor 6 (-0.10 <= x < -0.05)
SlogP_VSA6	SlogP_VSA	MOE logP VSA Descriptor 6 (0.15 <= x < 0.20)
SlogP_VSA10	SlogP_VSA	MOE logP VSA Descriptor 10 (0.40 <= x < 0.50)
VSA_EState1	VSA_Estate	VSA EState Descriptor 1 (-inf < x < 4.78)
VSA_EState7	VSA_Estate	VSA EState Descriptor 7 (6.07 <= x < 6.45)
Kappa3	Kappa descriptors	Hall-Kier Kappa2 value
SlogP_VSA4	SlogP_VSA	MOE logP VSA Descriptor 4 (0.00 <= x < 0.10)
VSA_EState5	VSA_Estate	VSA EState Descriptor 5 (5.74 <= x < 6.00)
EState_VSA9	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
PEOE_VSA10	PEOE_VSA	MOE Charge VSA Descriptor 10 (0.10 <= x < 0.15)
PEOE_VSA4	PEOE_VSA	MOE Charge VSA Descriptor 4 (-0.20 <= x < -0.15)
PEOE_VSA1	PEOE_VSA	MOE Charge VSA Descriptor 1 (-inf < x < -0.30)
SMR_VSA10	SMR_VSA	MOE MR VSA Descriptor 10 (4.00 <= x < inf)
EState_VSA7	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
FractionCSP3	FractionCSP3	The fraction of C atoms that are SP3 hybridized.
VSA_EState6	VSA_Estate	VSA EState Descriptor 6 (6.00 <= x < 6.07)
SlogP_VSA8	SlogP_VSA	MOE logP VSA Descriptor 8 (0.25 <= x < 0.30)
NumHDonors	NumHDonors	Number of Hydrogen Bond Donors
NumHeteroatoms	NumHeteroatoms	Number of Heteroatoms
SlogP_VSA12	SlogP_VSA	MOE logP VSA Descriptor 12 (0.60 <= x < inf)
MolLogP	MolLogP	Wildman-Crippen LogP value.Wildman and Crippen JCICS 39:868-73 (1999)
NumHAcceptors	NumHAcceptors	Number of Hydrogen Bond Acceptors
EState_VSA10	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
NumRotatableBonds	NumRotatableBonds	Number of Rotatable Bonds
EState_VSA5	EState_VSA	MOE-type descriptors using EState indices and surface area contributions (developed at RD, not described in the CCG paper).
RingCount	RingCount	The number of rings for a molecule
BertzCT	BertzCT	A topological index meant to quantify "complexity" of molecules.J. Am. Chem. Soc. 103:3599-601 (1981).
SMR_VSA6	SMR_VSA	MOE MR VSA Descriptor 6 (2.75 <= x < 3.05)
fr_COO2	fr_COO2	Number of carboxylic acids
fr_allylic_oxid	fr_allylic_oxid	Number of allylic oxidation sites excluding steroid dienone
Ipc	Ipc	the information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule.

Table S1(continued) The brief definitions of 87 selected descriptors

Descriptor	Type	Description
fr_aniline	fr_aniline	Number of anilines
NumAromaticRings	NumAromaticRings	The number of aromatic rings for a molecule
PEOE_VSA3	PEOE_VSA	MOE Charge VSA Descriptor 3 (-0.25 <= x < -0.20)
MinEStateIndex	Estate Index	Returns a tuple of EState indices for the molecule, Reference: Hall, Mohney and Kier. JCICS _31_ 76-81 (1991)
NOCount	NOCount	Number of Nitrogens and Oxygens
fr_NH0	fr_NH0	Number of Tertiary amines
fr_phenol_noOrthoHl	fr_phenol_noOrthoHbond	Number of phenolic OH excluding ortho intramolecular Hbond substituents
qed	qed	the quantitative estimation of drug-likeness

Table S2 Overview of the features employed for model development

Descriptor	Abbreviation	Original size	Selected size
Molecular Descriptors	MD	208	87
MACCS	MAC	167	71
Pubchem	Pub	881	245
KlekotaRoth	KR	4860	1103
CDK Extended	Ext	1024	503
Daylight	Day	1024	508
CDK GraphOnly	Gra	1024	405
Morgan (1024)	M1024	1024	512
Morgan (2048)	M2048	2048	1023

Table S3. The optimal hyperparameters of five algorithms

Algorithms	Hyperparameters
BNB	
XGBoost	n_estimators=100, learning_rate=0.1,gamma=0.5,max_depth=12,min_child_weight=7
RF	random_state=42, max_depth=10, n_estimators=200
SVM	gamma=0.01, C=1.0, kernel='rbf'
AdaBoost	learning_rate=0.9, n_estimators=200， random_state=42

Table S4 Five-fold stratified cross validation performances of combinatorial models for predicting hPXR activators

RDKit descriptors

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.771±0.02	0.623±0.02	0.616±0.04	0.316±0.03	0.417±0.03	0.290±0.04	0.316±0.04
RF	0.907±0.01	0.829±0.01	0.702±0.03	0.773±0.03	0.735±0.02	0.634±0.03	0.636±0.03
SVM	0.848±0.01	0.786±0.01	0.550±0.02	0.809±0.02	0.654±0.02	0.496±0.03	0.517±0.02
XGBoost	0.913±0.01	0.841±0.01	0.726±0.02	0.788±0.02	0.756±0.02	0.663±0.02	0.665±0.02
AdaBoost	0.895±0.01	0.797±0.02	0.736±0.03	0.682±0.03	0.708±0.02	0.608±0.03	0.610±0.03

Pubchem

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.766±0.02	0.650±0.02	0.561±0.04	0.416±0.04	0.477±0.03	0.326±0.04	0.332±0.04
RF	0.882±0.01	0.795±0.01	0.656±0.02	0.727±0.03	0.689±0.02	0.570±0.03	0.572±0.03
SVM	0.883±0.01	0.806±0.01	0.607±0.02	0.797±0.02	0.689±0.02	0.556±0.03	0.567±0.03
XGBoost	0.891±0.01	0.812±0.01	0.663±0.02	0.763±0.02	0.709±0.02	0.595±0.03	0.598±0.03
AdaBoost	0.872±0.01	0.755±0.02	0.700±0.03	0.603±0.03	0.646±0.02	0.533±0.03	0.536±0.03

CDK Extended

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.777±0.02	0.622±0.02	0.542±0.04	0.350±0.03	0.425±0.03	0.275±0.04	0.286±0.04
RF	0.87±0.01	0.787±0.02	0.629±0.02	0.727±0.03	0.674±0.02	0.546±0.03	0.549±0.03
SVM	0.881±0.01	0.811±0.01	0.612±0.02	0.804±0.02	0.694±0.02	0.564±0.02	0.575±0.02
XGBoost	0.884±0.01	0.808±0.01	0.652±0.02	0.762±0.02	0.702±0.02	0.584±0.03	0.588±0.03
AdaBoost	0.865±0.01	0.752±0.02	0.691±0.03	0.600±0.04	0.642±0.03	0.526±0.03	0.528±0.03

CDK GraphOnly

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.719±0.02	0.565±0.01	0.558±0.05	0.182±0.03	0.273±0.04	0.165±0.03	0.204±0.04
RF	0.803±0.02	0.727±0.02	0.485±0.03	0.735±0.03	0.584±0.02	0.391±0.03	0.410±0.03
SVM	0.802±0.01	0.733±0.02	0.459±0.02	0.808±0.03	0.585±0.02	0.375±0.03	0.412±0.03
XGBoost	0.818±0.02	0.733±0.02	0.497±0.02	0.732±0.03	0.592±0.02	0.405±0.03	0.422±0.03
AdaBoost	0.799±0.01	0.626±0.02	0.646±0.05	0.315±0.03	0.422±0.03	0.301±0.03	0.332±0.03

Morgan(2048)

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.818±0.01	0.681±0.02	0.661±0.03	0.443±0.04	0.531±0.03	0.405±0.03	0.417±0.03
RF	0.83±0.01	0.752±0.01	0.511±0.02	0.767±0.03	0.613±0.02	0.434±0.01	0.454±0.03
SVM	0.871±0.01	0.791±0.01	0.578±0.02	0.791±0.03	0.667±0.02	0.521±0.03	0.535±0.03
XGBoost	0.857±0.01	0.775±0.02	0.600±0.02	0.724±0.02	0.656±0.02	0.516±0.03	0.521±0.03
AdaBoost	0.847±0.01	0.699±0.02	0.699±0.04	0.471±0.04	0.562±0.03	0.444±0.04	0.459±0.04

MACCS

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.770±0.02	0.572±0.03	0.495±0.05	0.227±0.07	0.305±0.07	0.172±0.06	0.193±0.05
RF	0.874±0.01	0.789±0.01	0.684±0.02	0.692±0.03	0.688±0.02	0.575±0.03	0.576±0.03
SVM	0.858±0.01	0.785±0.01	0.569±0.02	0.783±0.03	0.659±0.02	0.509±0.02	0.523±0.02
XGBoost	0.875±0.01	0.800±0.01	0.649±0.02	0.744±0.03	0.693±0.02	0.572±0.02	0.575±0.02
AdaBoost	0.849±0.01	0.74±0.02	0.673±0.03	0.583±0.03	0.624±0.02	0.502±0.03	0.505±0.03

Klekota-Roth

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.789±0.02	0.657±0.02	0.644±0.03	0.392±0.04	0.486±0.03	0.359±0.04	0.377±0.03
RF	0.831±0.02	0.746±0.02	0.628±0.02	0.624±0.04	0.625±0.03	0.492±0.03	0.492±0.03
SVM	0.867±0.01	0.782±0.02	0.597±0.02	0.745±0.03	0.663±0.02	0.523±0.03	0.530±0.03
XGBoost	0.864±0.01	0.780±0.02	0.622±0.02	0.716±0.03	0.665±0.02	0.533±0.03	0.536±0.03
AdaBoost	0.848±0.01	0.726±0.01	0.722±0.03	0.525±0.03	0.608±0.02	0.496±0.03	0.507±0.03

Daylight

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.765±0.02	0.614±0.02	0.542±0.04	0.328±0.03	0.408±0.03	0.261±0.04	0.275±0.04
RF	0.852±0.01	0.761±0.01	0.598±0.02	0.692±0.03	0.640±0.02	0.497±0.03	0.500±0.03
SVM	0.872±0.01	0.794±0.01	0.586±0.02	0.789±0.03	0.672±0.02	0.530±0.03	0.542±0.03
XGBoost	0.877±0.01	0.799±0.01	0.642±0.02	0.747±0.03	0.690±0.02	0.568±0.02	0.571±0.02
AdaBoost	0.861±0.01	0.742±0.02	0.702±0.03	0.571±0.04	0.629±0.03	0.514±0.03	0.520±0.03

Morgan(1024)

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.806±0.02	0.666±0.02	0.638±0.03	0.418±0.04	0.504±0.03	0.373±0.03	0.388±0.03
RF	0.834±0.01	0.757±0.01	0.527±0.02	0.757±0.03	0.621±0.02	0.450±0.03	0.467±0.03
SVM	0.861±0.01	0.784±0.01	0.564±0.02	0.787±0.02	0.657±0.02	0.504±0.03	0.520±0.03
XGBoost	0.859±0.01	0.776±0.02	0.601±0.02	0.724±0.03	0.657±0.02	0.518±0.03	0.522±0.03
AdaBoost	0.848±0.01	0.710±0.02	0.686±0.03	0.502±0.04	0.579±0.03	0.458±0.04	0.468±0.04

Table S5 External validation performances of combinatorial models for predicting hPXR activators

RDKit descriptors

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.649	0.649	0.627	0.380	0.473	0.341	0.359
RF	0.849	0.849	0.691	0.832	0.755	0.656	0.661
SVM	0.800	0.800	0.547	0.854	0.667	0.508	0.537
XGBoost	0.860	0.860	0.728	0.832	0.777	0.689	0.692
AdaBoost	0.813	0.813	0.728	0.723	0.725	0.627	0.627

MACCS

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.656	0.656	0.537	0.453	0.491	0.329	0.331
RF	0.825	0.825	0.682	0.781	0.728	0.621	0.624
SVM	0.799	0.799	0.566	0.825	0.672	0.522	0.542
XGBoost	0.806	0.806	0.650	0.759	0.700	0.581	0.585
AdaBoost	0.758	0.758	0.680	0.620	0.649	0.530	0.531

Pubchem

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.720	0.720	0.588	0.588	0.588	0.439	0.439
RF	0.812	0.812	0.638	0.785	0.704	0.582	0.588
SVM	0.822	0.822	0.596	0.850	0.701	0.566	0.585
XGBoost	0.845	0.845	0.674	0.836	0.746	0.641	0.648
AdaBoost	0.782	0.782	0.713	0.661	0.686	0.578	0.579

Klekota-Roth

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.644	0.644	0.620	0.369	0.462	0.330	0.348
RF	0.763	0.763	0.614	0.679	0.645	0.508	0.509
SVM	0.808	0.808	0.614	0.796	0.693	0.563	0.572
XGBoost	0.795	0.795	0.624	0.752	0.682	0.553	0.558
AdaBoost	0.746	0.746	0.754	0.558	0.642	0.537	0.548

CDK Extended

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.659	0.659	0.534	0.464	0.496	0.332	0.333
RF	0.814	0.814	0.628	0.799	0.703	0.578	0.587
SVM	0.827	0.827	0.606	0.854	0.709	0.579	0.597
XGBoost	0.829	0.829	0.643	0.821	0.721	0.604	0.613
AdaBoost	0.769	0.769	0.655	0.664	0.659	0.536	0.536

Daylight

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.621	0.621	0.536	0.350	0.424	0.272	0.282
RF	0.802	0.802	0.607	0.788	0.686	0.552	0.562
SVM	0.829	0.829	0.611	0.854	0.712	0.584	0.602
XGBoost	0.810	0.810	0.624	0.792	0.698	0.571	0.579
AdaBoost	0.749	0.749	0.669	0.606	0.636	0.514	0.515

CDK GraphOnly

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.566	0.566	0.542	0.190	0.281	0.167	0.201
RF	0.722	0.722	0.463	0.759	0.575	0.368	0.394
SVM	0.737	0.737	0.457	0.828	0.589	0.376	0.419
XGBoost	0.748	0.748	0.495	0.781	0.606	0.418	0.443
AdaBoost	0.645	0.645	0.647	0.361	0.464	0.338	0.361

Morgan(1024)

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.712	0.712	0.644	0.529	0.581	0.450	0.454
RF	0.777	0.777	0.516	0.836	0.638	0.462	0.494
SVM	0.801	0.801	0.568	0.828	0.674	0.524	0.545
XGBoost	0.804	0.804	0.610	0.788	0.688	0.555	0.565
AdaBoost	0.727	0.727	0.709	0.533	0.608	0.493	0.502

Morgan(2048)

	AUC	BA	Precision	Recall	F1-score	CK	MCC
BNB	0.691	0.691	0.655	0.471	0.548	0.419	0.429
RF	0.774	0.774	0.499	0.858	0.631	0.446	0.486
SVM	0.803	0.803	0.572	0.828	0.677	0.529	0.549
XGBoost	0.805	0.805	0.606	0.796	0.688	0.554	0.565
AdaBoost	0.722	0.722	0.708	0.522	0.601	0.485	0.495

Table S6 Performance of in domain and out of domain chemicals in the external test set for the top ten combinatorial classification models

Application domain analysis based on Tanimoto similarity

	ID							OD						
Models	BA	Precision	Recall	F1	AUC	CK	MCC	BA	Precision	Recall	F1	AUC	CK	MCC
RDKitMD-XGBoost	0.868	0.737	0.833	0.782	0.868	0.705	0.707	0.851	0.72	0.831	0.771	0.851	0.671	0.675
RDKitMD-RF	0.841	0.685	0.804	0.74	0.841	0.645	0.649	0.856	0.696	0.86	0.77	0.856	0.664	0.672
Pub-XGBoost	0.855	0.669	0.848	0.748	0.855	0.651	0.66	0.835	0.679	0.824	0.744	0.835	0.629	0.635
Ext-XGBoost	0.855	0.678	0.841	0.751	0.855	0.657	0.664	0.799	0.609	0.801	0.692	0.799	0.545	0.557
Day-SVM	0.841	0.634	0.841	0.723	0.841	0.614	0.626	0.814	0.59	0.868	0.702	0.814	0.55	0.575
Ext-SVM	0.854	0.647	0.862	0.739	0.854	0.636	0.649	0.796	0.569	0.846	0.68	0.796	0.516	0.54
MAC-RF	0.851	0.687	0.826	0.75	0.851	0.657	0.663	0.798	0.676	0.735	0.704	0.798	0.58	0.581
Pub-SVM	0.835	0.609	0.848	0.709	0.835	0.592	0.608	0.805	0.583	0.853	0.693	0.805	0.536	0.559
Ext-RF	0.84	0.681	0.804	0.738	0.84	0.642	0.646	0.783	0.581	0.794	0.671	0.783	0.51	0.524
RDKitMD - AdaBoost	0.818	0.744	0.717	0.731	0.818	0.644	0.644	0.806	0.712	0.728	0.72	0.806	0.607	0.607

Application domain analysis based on Euclidean distance

	ID							OD						
Models	BA	Precision	Recall	F1	AUC	CK	MCC	BA	Precision	Recall	F1	AUC	CK	MCC
RDKitMD-XGBoost	0.869	0.758	0.802	0.779	0.869	0.722	0.722	0.832	0.708	0.856	0.775	0.832	0.634	0.642
RDKitMD-RF	0.852	0.729	0.777	0.752	0.852	0.687	0.688	0.819	0.667	0.876	0.757	0.819	0.595	0.611
Pub-XGBoost	0.862	0.688	0.818	0.747	0.862	0.677	0.681	0.808	0.663	0.850	0.745	0.808	0.578	0.591
Ext-XGBoost	0.864	0.680	0.826	0.746	0.864	0.674	0.680	0.770	0.616	0.817	0.702	0.770	0.503	0.517
Day-SVM	0.852	0.643	0.818	0.720	0.852	0.639	0.646	0.775	0.590	0.882	0.707	0.775	0.493	0.525
Ext-SVM	0.864	0.650	0.843	0.734	0.864	0.656	0.665	0.759	0.576	0.863	0.691	0.759	0.466	0.496
MAC-RF	0.864	0.697	0.818	0.753	0.864	0.684	0.688	0.775	0.669	0.752	0.708	0.775	0.534	0.537
Pub-SVM	0.845	0.628	0.810	0.708	0.845	0.622	0.630	0.764	0.574	0.882	0.696	0.764	0.470	0.506
Ext-RF	0.838	0.674	0.769	0.718	0.838	0.641	0.644	0.761	0.597	0.824	0.692	0.761	0.480	0.499
RDKitMD - AdaBoost	0.801	0.741	0.661	0.699	0.801	0.628	0.63	0.804	0.72	0.771	0.744	0.804	0.598	0.599

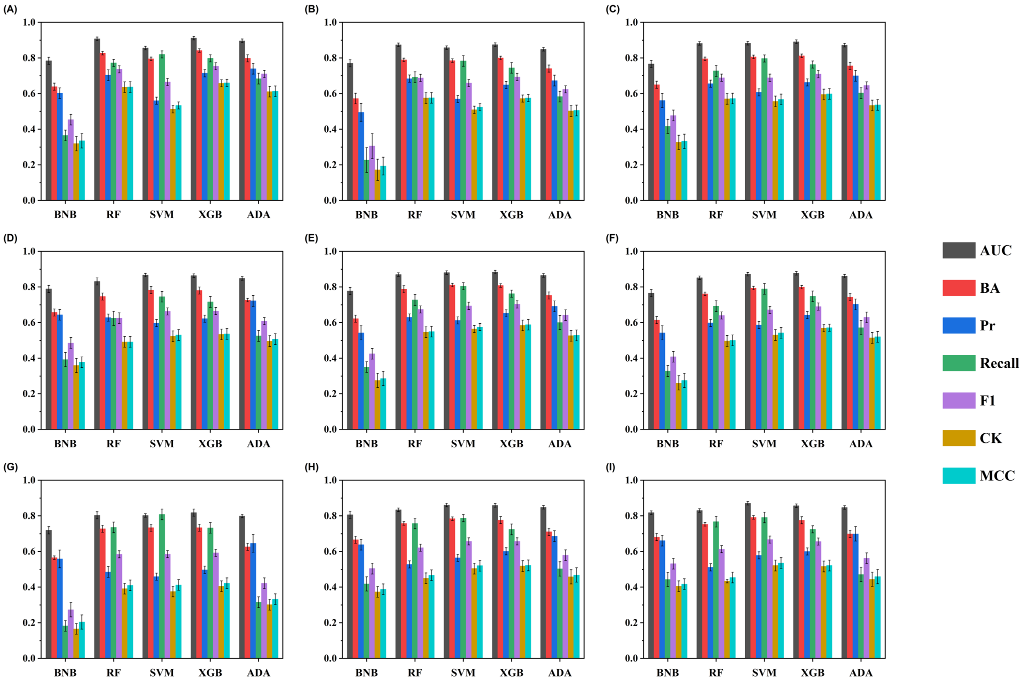

Figure S1. Five-fold repeated stratified cross validation performances of individual models constructed by five machine learning algorithms and nine molecular features. The sub-figures show the results using nine molecular features. The y-axis gives the performance values and different metrics are depicted by colors. Five machine learning algorithms are grouped and labeled at the x-axis. (A) RDKit molecular descriptors (B) MACCS fingerprint (C) Pubchem fingerprint (D) KlekotaRoth fingerprint (E) CDK Extended fingerprint (F) Daylight fingerprint (G) CDK GraphOnly fingerprint (H) Morgan (1024) fingerprint (I) Morgan (2048) fingerprint