Supporting Information

ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches

Zhenxing Wu[a], Tailong Lei[a], Chao Shen[a], Zhe Wang[a], Dongsheng Cao[c], Tingjun Hou[a,b,*]

[a]Hangzhou Institute of Innovative Medicine, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China; [b]State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China; [c]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410004, Hunan, P. R. China

Corresponding author:

Tingjun Hou

***E-mail**: tingjunhou@zju.edu.cn

Table S1. The main hyperparameters for the RF models

hyperparameter	CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
n_estimators	1900	1700	1800	900	1900
min_samples_split	2	2	2	2	2
min_samples_leaf	2	8	8	10	10
max_depth	40	20	20	10	10
max_features	auto	auto	auto	auto	auto

Table S2. The main hyperparameters for the GBDT models

hyperparameter	CYP1A2	CYP2C9	CYP2C19	CYP2D6	CYP3A4
learning_rate	0.005	0.02	0.005	0.005	0.005
n_estimators	4800	350	1800	3200	4800
max_depth	11	13	7	15	5
min_samples_split	100	300	500	500	1100
min_samples_leaf	20	20	80	20	40
max_features	46	44	44	44	34
subsample	0.85	0.65	0.9	0.8	0.9

Table S3. Performances of different models on the training and test sets

RF	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.877	0.802	0.923	0.736	0.945	0.967	0.879	0.987	0.889	0.991
2C9	0.846	0.610	0.934	0.591	0.904	0.899	0.275	0.971	0.333	0.801
2C19	0.824	0.811	0.833	0.643	0.895	0.814	0.556	0.879	0.428	0.815
2D6	0.896	0.271	0.993	0.445	0.861	0.918	0.286	0.991	0.443	0.834
3A4	0.828	0.688	0.899	0.607	0.901	0.859	0.477	0.970	0.553	0.907
GBDT	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.900	0.849	0.930	0.785	0.960	0.974	0.907	0.990	0.913	0.991
2C9	0.860	0.683	0.927	0.635	0.919	0.899	0.391	0.958	0.397	0.821
2C19	0.837	0.829	0.844	0.671	0.906	0.813	0.634	0.858	0.460	0.825
2D6	0.913	0.457	0.983	0.568	0.892	0.918	0.442	0.973	0.496	0.863
3A4	0.851	0.739	0.908	0.661	0.923	0.879	0.538	0.978	0.626	0.927
XGBoost	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.905	0.862	0.930	0.796	0.962	0.974	0.916	0.987	0.913	0.991
2C9	0.876	0.733	0.930	0.680	0.931	0.902	0.290	0.973	0.354	0.814
2C19	0.843	0.829	0.854	0.683	0.910	0.816	0.613	0.866	0.456	0.836
2D6	0.909	0.434	0.982	0.542	0.877	0.928	0.416	0.987	0.537	0.863
3A4	0.860	0.757	0.913	0.683	0.931	0.894	0.618	0.975	0.677	0.935
DNN	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.866	0.907	0.842	0.731	0.942	0.908	0.836	0.924	0.715	0.949
2C9	0.890	0.735	0.948	0.714	0.953	0.849	0.367	0.904	0.251	0.750
2C19	0.840	0.840	0.840	0.678	0.900	0.785	0.488	0.859	0.340	0.784
2D6	0.930	0.717	0.963	0.694	0.931	0.904	0.392	0.962	0.411	0.831
3A4	0.839	0.811	0.853	0.650	0.913	0.861	0.627	0.929	0.584	0.891

CNN	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.888	0.844	0.914	0.760	0.950	0.959	0.885	0.976	0.844	0.976
2C9	0.844	0.762	0.875	0.619	0.900	0.866	0.223	0.940	0.187	0.681
2C19	0.828	0.836	0.822	0.654	0.900	0.765	0.567	0.815	0.348	0.781
2D6	0.904	0.372	0.987	0.509	0.867	0.919	0.401	0.978	0.482	0.844
3A4	0.833	0.819	0.843	0.640	0.911	0.887	0.674	0.949	0.661	0.914

Table S4. Performances of the XGBoost models based on different sets of descriptors

PubFP	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.888	0.833	0.921	0.759	0.952	0.962	0.925	0.971	0.877	0.982
2C9	0.847	0.661	0.917	0.602	0.903	0.857	0.333	0.918	0.246	0.760
2C19	0.818	0.807	0.827	0.633	0.890	0.805	0.542	0.870	0.403	0.794
2D6	0.907	0.451	0.977	0.537	0.875	0.926	0.377	0.990	0.520	0.861
3A4	0.842	0.726	0.902	0.642	0.914	0.870	0.486	0.982	0.595	0.922
MorFP	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.881	0.823	0.916	0.745	0.948	0.969	0.897	0.985	0.896	0.989
2C9	0.846	0.652	0.918	0.597	0.902	0.887	0.275	0.958	0.286	0.785
2C19	0.816	0.803	0.826	0.628	0.895	0.762	0.451	0.840	0.282	0.778
2D6	0.906	0.423	0.981	0.527	0.879	0.929	0.403	0.990	0.543	0.780
3A4	0.854	0.731	0.918	0.669	0.918	0.856	0.410	0.986	0.543	0.905
KleFP	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.882	0.825	0.917	0.748	0.948	0.961	0.888	0.977	0.868	0.984
2C9	0.850	0.659	0.921	0.607	0.908	0.856	0.261	0.924	0.193	0.778
2C19	0.834	0.824	0.843	0.665	0.906	0.810	0.472	0.895	0.382	0.828
2D6	0.909	0.475	0.976	0.551	0.880	0.932	0.377	0.996	0.559	0.827
3A4	0.850	0.722	0.915	0.657	0.917	0.867	0.508	0.972	0.585	0.885
GraFP	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.851	0.774	0.898	0.680	0.923	0.928	0.776	0.962	0.755	0.970
2C9	0.834	0.613	0.916	0.562	0.885	0.872	0.261	0.943	0.232	0.764
2C19	0.793	0.771	0.811	0.582	0.872	0.769	0.444	0.851	0.290	0.759
2D6	0.899	0.384	0.979	0.484	0.852	0.909	0.312	0.978	0.396	0.858
3A4	0.826	0.673	0.904	0.601	0.896	0.851	0.426	0.975	0.525	0.901

MOE	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.898	0.850	0.926	0.781	0.958	0.967	0.907	0.981	0.891	0.986
2C9	0.872	0.727	0.926	0.670	0.927	0.871	0.377	0.928	0.305	0.792
2C19	0.840	0.841	0.839	0.671	0.910	0.790	0.577	0.843	0.394	0.807
2D6	0.913	0.443	0.986	0.567	0.897	0.922	0.429	0.979	0.511	0.849
3A4	0.856	0.771	0.890	0.677	0.925	0.877	0.585	0.963	0.623	0.929
PaDel	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.904	0.858	0.931	0.793	0.961	0.973	0.907	0.987	0.907	0.992
2C9	0.866	0.716	0.922	0.654	0.921	0.904	0.188	0.987	0.305	0.837
2C19	0.849	0.844	0.853	0.695	0.918	0.810	0.662	0.847	0.467	0.843
2D6	0.912	0.445	0.984	0.561	0.885	0.912	0.403	0.970	0.449	0.822
3A4	0.858	0.752	0.913	0.679	0.930	0.893	0.611	0.975	0.673	0.934
Pub+MOE	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.901	0.853	0.929	0.787	0.961	0.969	0.907	0.983	0.896	0.987
2C9	0.847	0.661	0.917	0.602	0.903	0.857	0.333	0.921	0.278	0.786
2C19	0.853	0.852	0.853	0.703	0.922	0.789	0.634	0.828	0.418	0.829
2D6	0.914	0.466	0.984	0.576	0.893	0.932	0.494	0.982	0.579	0.864
3A4	0.861	0.769	0.909	0.687	0.931	0.877	0.562	0.969	0.619	0.935
Pub+PaDel	Training set					Test set				
	ACC	SE	SP	MCC	AUC	ACC	SE	SP	MCC	AUC
1A2	0.905	0.862	0.930	0.796	0.962	0.974	0.916	0.987	0.913	0.991
2C9	0.876	0.733	0.930	0.680	0.931	0.902	0.290	0.973	0.354	0.814
2C19	0.850	0.843	0.856	0.697	0.917	0.823	0.669	0.861	0.493	0.842
2D6	0.909	0.434	0.982	0.542	0.877	0.928	0.416	0.987	0.537	0.863
3A4	0.860	0.757	0.913	0.683	0.931	0.894	0.618	0.975	0.677	0.935

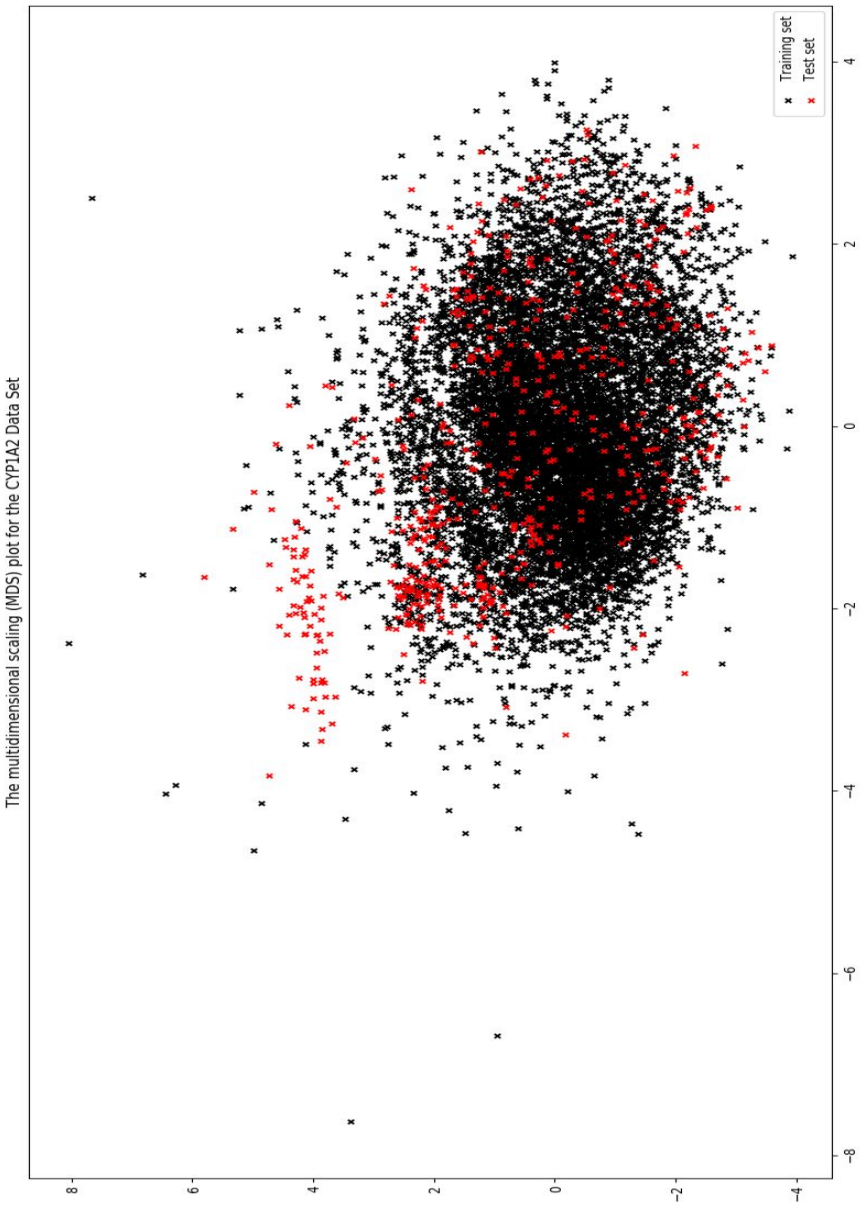
Table S5. The descriptions of the representative molecular descriptors

Descriptor	Description
nAcid	Number of acidic groups
ALogP	Ghose-Crippen LogKow
ATS8m	Broto-Moreau autocorrelation - lag 8 / weighted by mass
ATS4e	Broto-Moreau autocorrelation - lag 4 / weighted by Sanderson electronegativities
ATS5i	Broto-Moreau autocorrelation - lag 5 / weighted by first ionization potential
AATS4m	Average Broto-Moreau autocorrelation - lag 4 / weighted by mass
AATS8v	Average Broto-Moreau autocorrelation - lag 8 / weighted by van der Waals volumes
AATS0p	Average Broto-Moreau autocorrelation - lag 0 / weighted by polarizabilities
AATS5p	Average Broto-Moreau autocorrelation - lag 5 / weighted by polarizabilities
AATS6p	Average Broto-Moreau autocorrelation - lag 6 / weighted by polarizabilities
AATS4i	Average Broto-Moreau autocorrelation - lag 4 / weighted by first ionization potential
AATS0s	Average Broto-Moreau autocorrelation - lag 0 / weighted by I-state
ATSC5s	Centered Broto-Moreau autocorrelation - lag 5 / weighted by I-state
GATS1m	Geary autocorrelation - lag 1 / weighted by mass
GATS5m	Geary autocorrelation - lag 5 / weighted by mass
GATS1e	Geary autocorrelation - lag 1 / weighted by Sanderson electronegativities
GATS1i	Geary autocorrelation - lag 1 / weighted by first ionization potential
GATS2i	Geary autocorrelation - lag 2 / weighted by first ionization potential
GATS5i	Geary autocorrelation - lag 5 / weighted by first ionization potential
nBase	Number of basic groups.
BCUTp-1h	nlow highest polarizability weighted BCUTS
SpMax3_Bhm	Largest absolute eigenvalue of Burden modified matrix - n 3 / weighted by relative mass

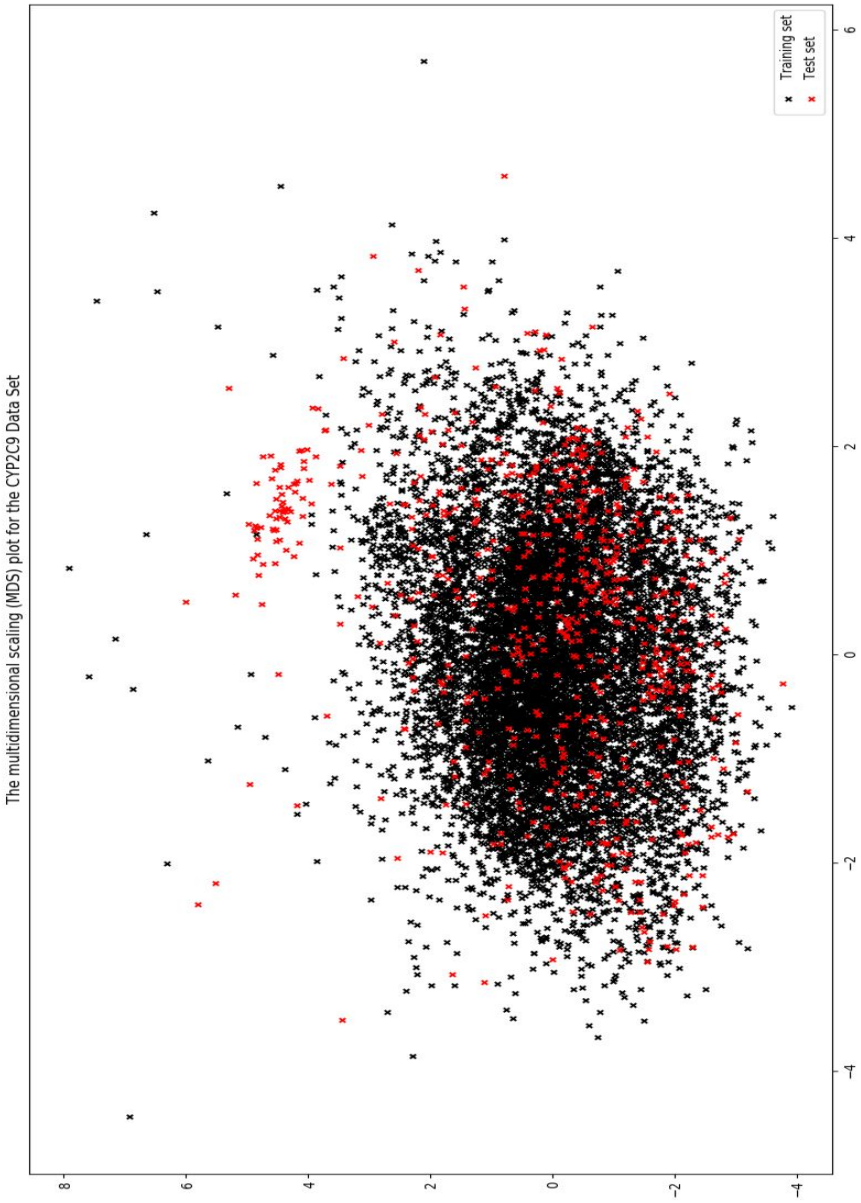
SpMin2_Bhs	Smallest absolute eigenvalue of Burden modified matrix - n 2 / weighted by relative I-state
C2SP2	Doubly bound carbon bound to two other carbons
ASP-7	Average simple path, order 7
Mp	Mean atomic polarizabilities (scaled on carbon atom)
CrippenLogP	Crippen's logP
nwHBa	Count of E-States for weak Hydrogen Bond acceptors
SwHBa	Sum of E-States for weak hydrogen bond acceptors
SHBint2	Sum of E-State descriptors of strength for potential hydrogen bonds of path length 2
SHBint7	Sum of E-State descriptors of strength for potential hydrogen bonds of path length 7
SHsOH	Sum of atom-type H E-State: -OH
SaaCH	Sum of atom-type E-State: :CH:
SsssCH	Sum of atom-type E-State: >CH-
SdssC	Sum of atom-type E-State: =C<
SaasC	Sum of atom-type E-State: :C:-
SaaN	Sum of atom-type E-State: :N:
SddsN	Sum of atom-type E-State: -N<<
minHBa	Minimum E-States for (strong) Hydrogen Bond acceptors
minHBint2	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 2
minHBint10	Minimum E-State descriptors of strength for potential Hydrogen Bonds of path length 10
minHsOH	Minimum atom-type H E-State: -OH
mindssC	Minimum atom-type E-State: =C<
minaaN	Minimum atom-type E-State: :N:
minsssN	Minimum atom-type E-State: >N-

minddsN	Minimum atom-type E-State: -N<<
mindO	Minimum atom-type E-State: =O
maxwHBa	Maximum E-States for weak Hydrogen Bond acceptors
maxHBint2	Maximum E-State descriptors of strength for potential Hydrogen Bonds of path length 2
maxHaaCH	Maximum atom-type H E-State: :CH:
maxaaCH	Maximum atom-type E-State: :CH:
maxaasC	Maximum atom-type E-State: :C:-
maxssNH	Maximum atom-type E-State: -NH-
LipoaffinityIndex	Lipoaffinity index
ETA_Beta	A measure of electronic features of the molecule
ETA_Beta_ns	A measure of electron-richness of the molecule
ETA_BetaP_ns	A measure of electron-richness of the molecule relative to molecular size
ETA_BetaP_ns_d	A measure of lone electrons entering into resonance relative to molecular size
IC3	Information content index (neighborhood symmetry of 3-order)
IC4	Information content index (neighborhood symmetry of 4-order)
IC5	Information content index (neighborhood symmetry of 5-order)
nAtomP	Number of atoms in the largest pi system
nAtomLAC	Number of atoms in the longest aliphatic chain
MDEO-11	Molecular distance edge between all primary oxygens
MLFER_BH	Overall or summation solute hydrogen bond basicity
MLFER_S	Combined dipolarity/polarizability
R_TpiPCTPC	Ratio of total conventional bond order (up to order 10) with total path count (up to order 10)
topoRadius	Topological radius (minimum atom eccentricity)
JGI10	Mean topological charge index of order 10
TopoPSA	Topological polar surface area

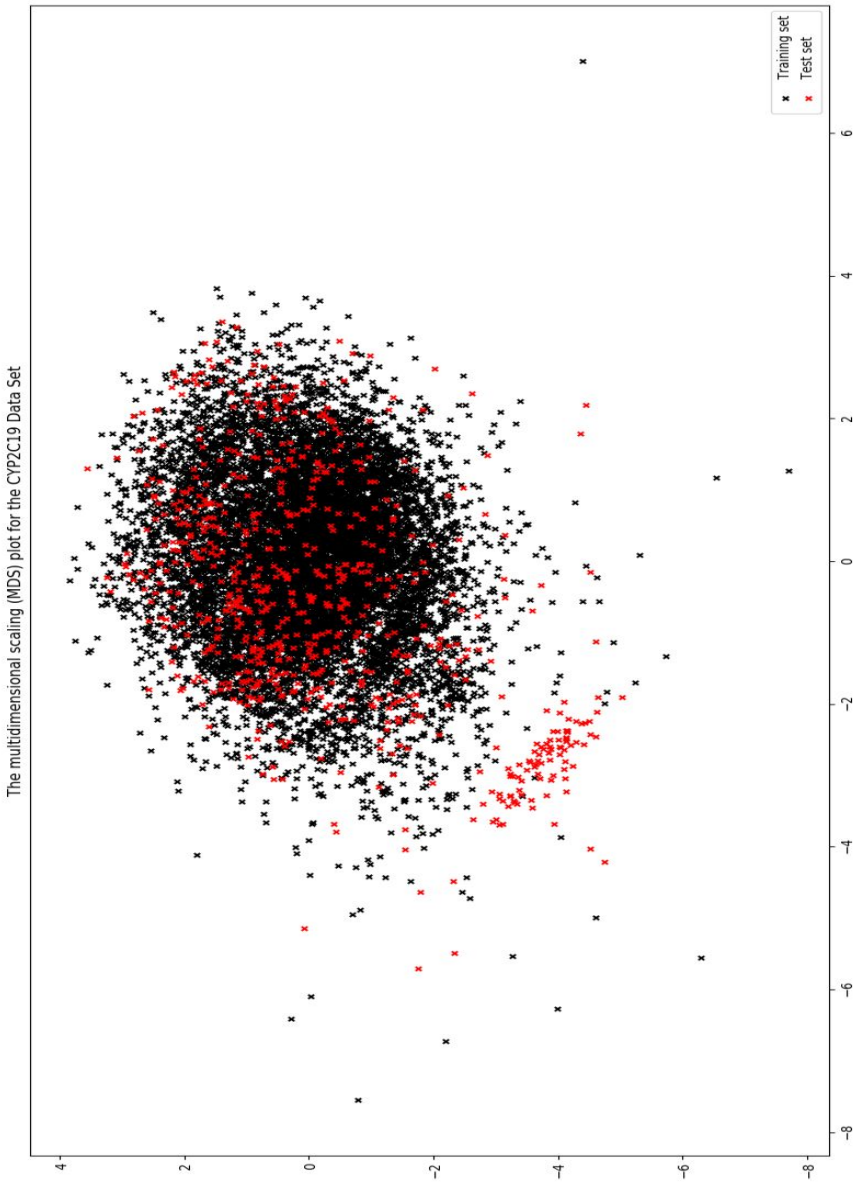
WTPT-5	Sum of path lengths starting from nitrogens
XLogP	XLogP
PubchemFP367	C(~H)(~O)(~O)
PubchemFP372	C(~H)(:C)(:N)
PubchemFP594	C-O-C-C=C


The multidimensional scaling (MDS) plot for the CYP1A2 Data Set
Training set
Test set
8
6
4
2
0
−2
−4
−8
−6
−4
−2
0
2
4

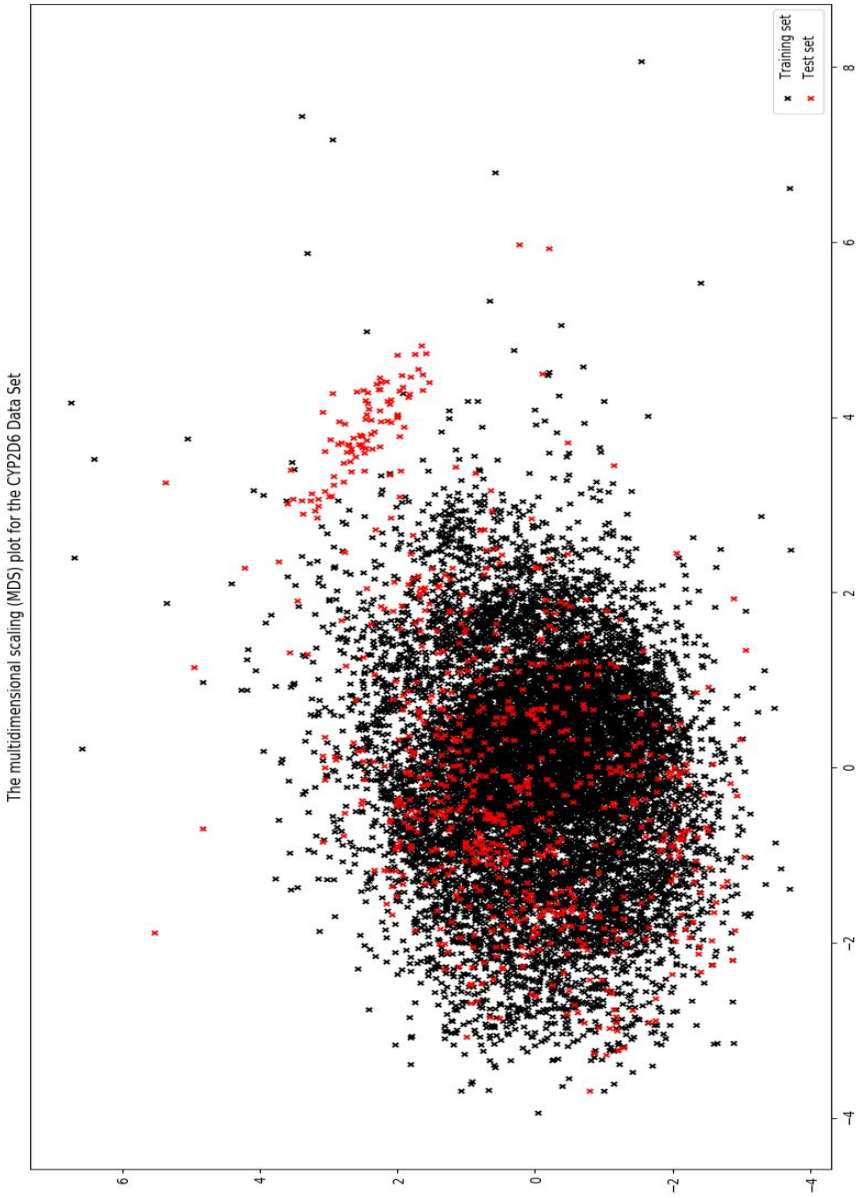
(A)


The multidimensional scaling (MDS) plot for the CYP2C9 Data Set
Training set
Test set
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6

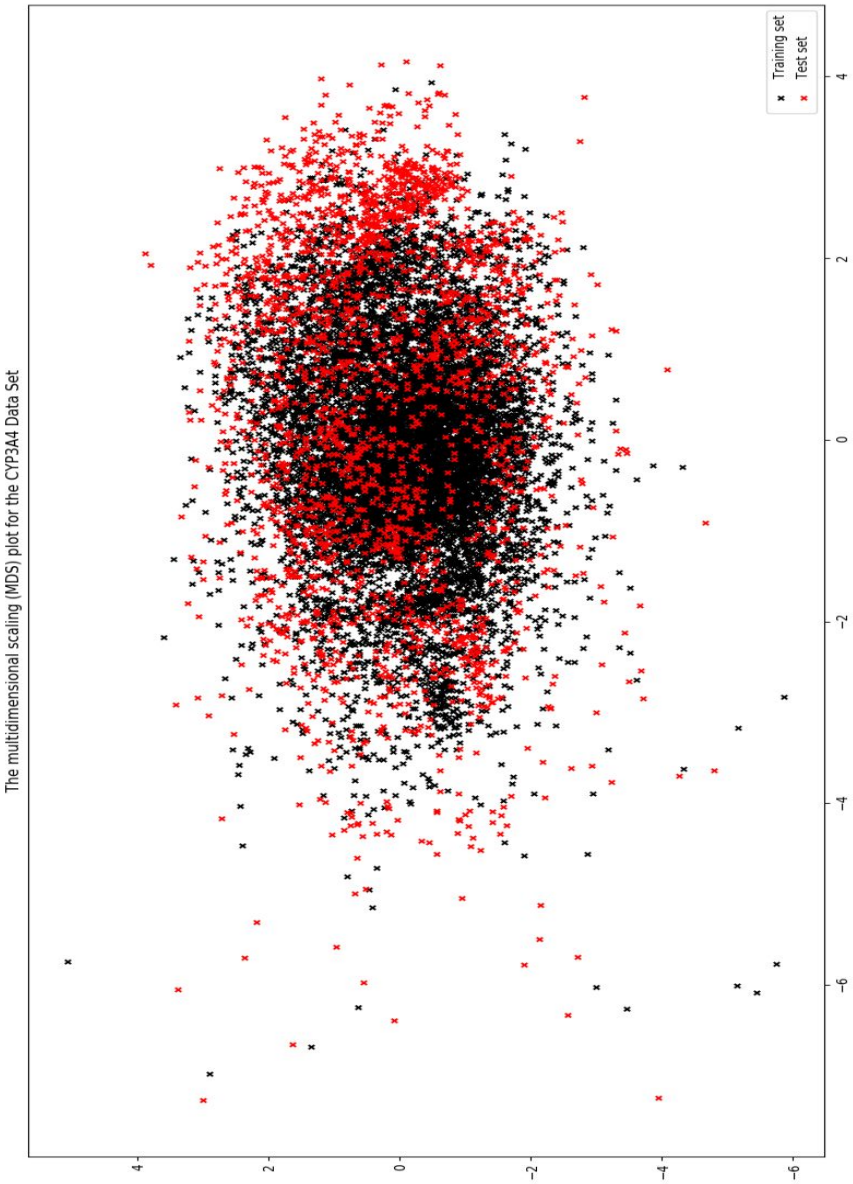
(B)


The multidimensional scaling (MDS) plot for the CYP2C19 Data Set
Training set
Test set
4
2
0
−2
−4
−6
−8
6
4
2
0
−2
−4
−6
−8

(C)

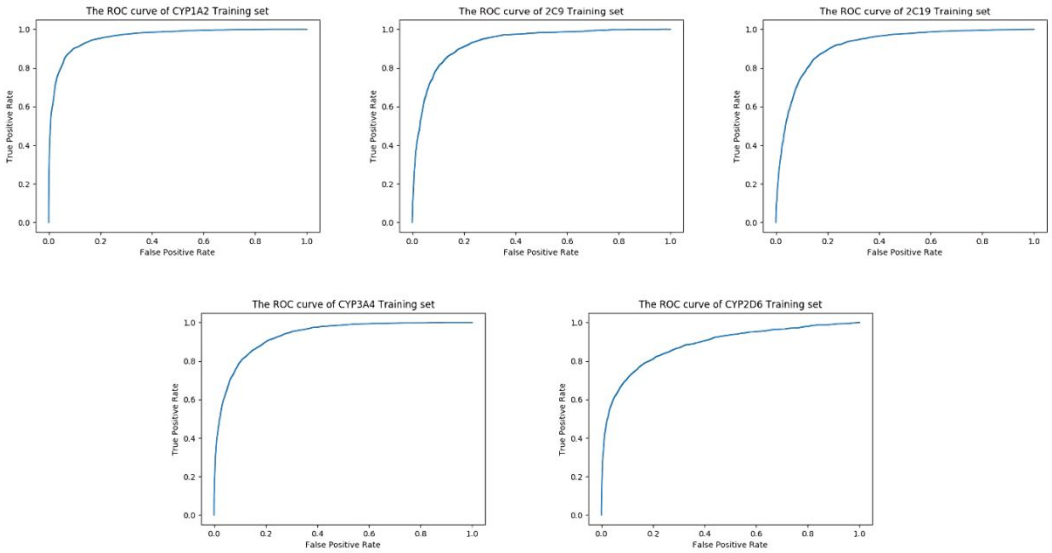

The multidimensional scaling (MDS) plot for the CYP2D6 Data Set
Training set
Test set
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8

(D)

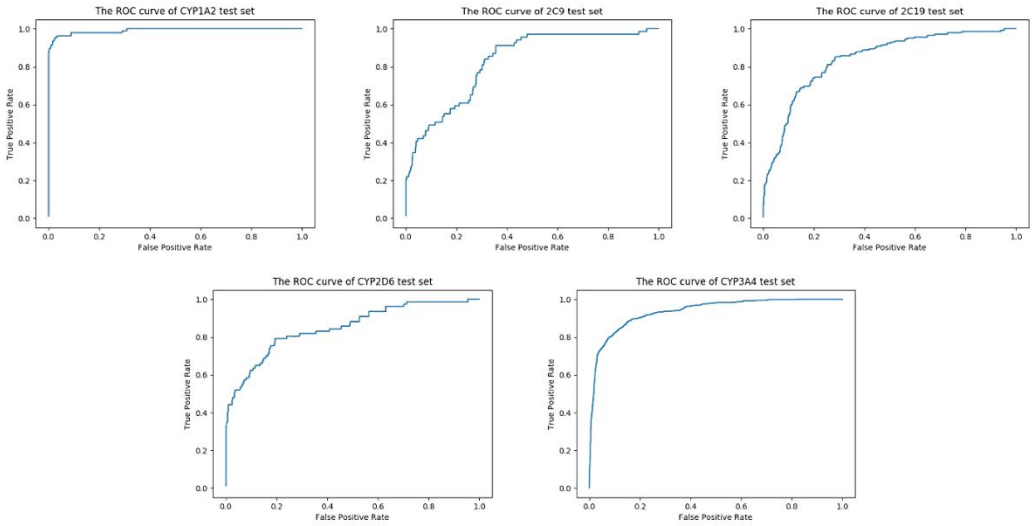


(E)

Figure S1. The multidimensional scaling (MDS) plots for the (A) CYP1A2, (B) CYP2C9, (C) CYP2C19, (D) CYP2D6, and (E) CYP3A4 datasets.



(A)



(B)

Figure S2. The ROC curves of different XGBoost models for (A) the training sets and (B) the test sets.