Application of machine learning methods for prediction of compound activities and SAR analysis

Aixia YAN, PhD
Department of Pharmaceutical Engineering,
Beijing University of Chemical Technology,
Beijing 100029, P. R. China

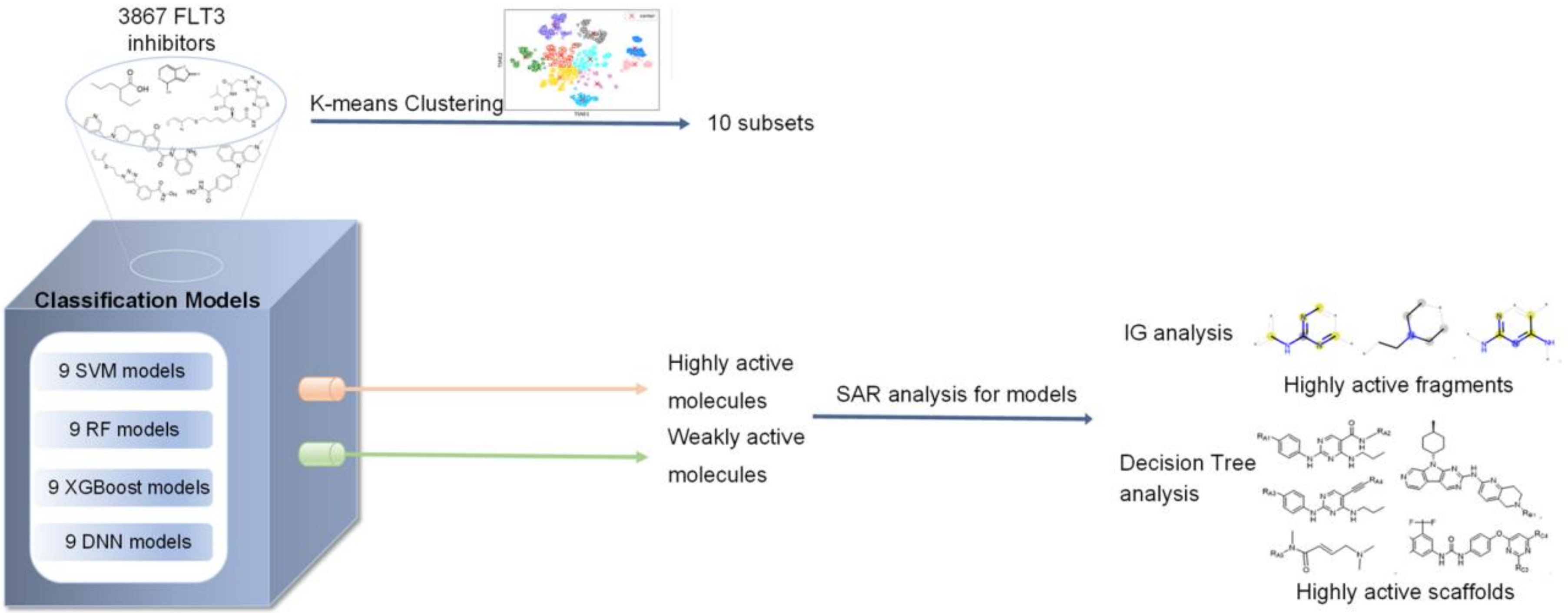
E-mail: yanax@mail.buct.edu.cn
http://www.cadd408.com/
ResearchGate: http://www.researchgate.net/profile/Aixia_Yan/

Prediction of bioactivities of small molecules using machine learning methods

- Classification of FLT3 inhibitors and SAR analysis by machine learning methods
- Classification of non-covalent Bruton’s tyrosine kinase (BTK) inhibitors and SAR analysis by machine learning methods
- A SAR and QSAR study on cyclin dependent kinase 4 (CDK4) inhibitors using machine learning methods

Classification of FLT3 inhibitors and SAR analysis by machine learning methods

We conducted 36 classification models based on support vector machine (SVM), random forest (RF), eXtreme Gradient Boosting (XGBoost), and deep neural networks (DNN) algorithms. The model built by deep neural networks (DNN) and TT fingerprints performed best on the test set with the highest prediction accuracy of 85.83% and Matthews correlation coefficient (MCC) of 0.72 and also performed well on the external test set. In addition, we clustered 3867 inhibitors into 11 subsets by the K-Means algorithm to figure out the structural characteristics of the reported FLT3.



Zhao, Y., et al. Yan, A.* Classification of FLT3 inhibitors and SAR analysis by machine learning methods. *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10640-8

Classification of FLT3 inhibitors and SAR analysis by machine learning methods

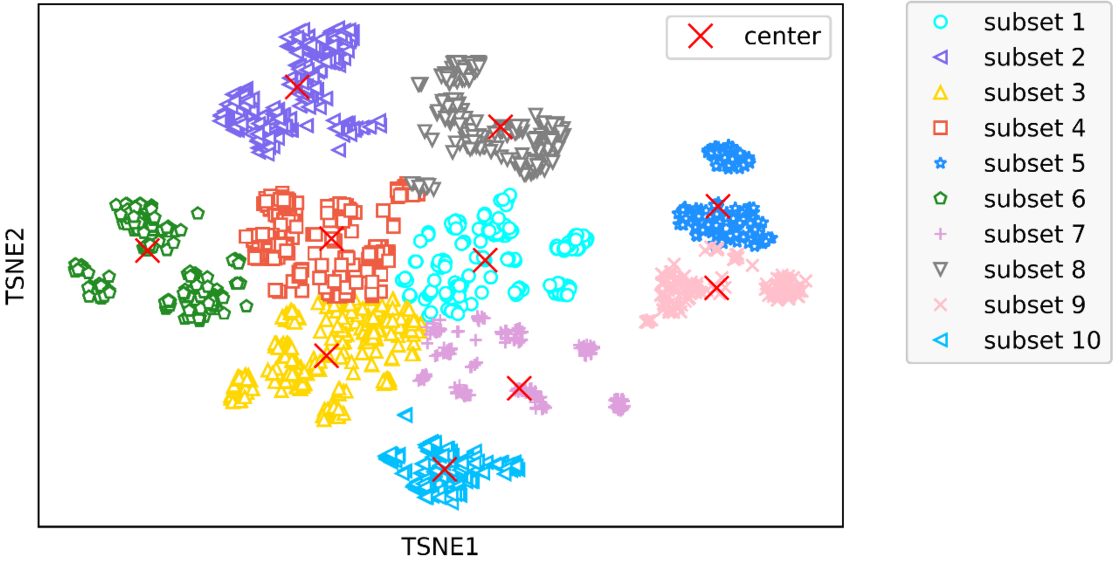
We have constructed 36 classification models for discriminating highly and weakly active inhibitors of FLT3. The accuracy rates of all the models were above 80%, and the highest accuracy reached 86% on the test set.

Model	Algorithm	Descriptors	Training set	Test set		External test set	
			5-CV	Q	MCC	Q	MCC
1A-1C	**SVM**	**MACCS**	81.36±0.275	0.828±0.006	0.66±0.015	67.29±7.78	0.36±0.148
1D-1E	**RF**	**MACCS**	81.93±0.110	0.830±0.004	0.66±0.006	60.95±8.96	0.22±0.167
1G-1I	**XGBoost**	**MACCS**	81.56±0.058	0.827±0.005	0.65±0.012	57.53±7.78	0.15±0.151
1J-1L	**DNN**	**MACCS**	81.89±0.227	0.833±0.008	0.66±0.015	54.93±2.90	0.10±0.045
2A-2C	**SVM**	**ECFP4**	84.83±0.390	0.856±0.002	0.71±0.006	80.06±3.25	0.61±0.059
2D-2E	**RF**	**ECFP4**	83.42±0.335	0.852±0.005	0.71±0.012	79.02±3.90	0.58±0.081
2G-2I	**XGBoost**	**ECFP4**	83.97±0.215	0.855±0.002	0.71±0.006	71.96±7.04	0.44±0.139
2J-2L	**DNN**	**ECFP4**	84.19±0.569	0.862±0.003	0.72±0.006	70.72±5.72	0.43±0.115
3A-3C	**SVM**	**TT**	84.20±0.367	0.852±0.002	0.70±0.006	84.63±6.55	0.71±0.109
3D-3E	**RF**	**TT**	83.80±0.546	0.851±0.005	0.70±0.012	82.76±1.77	0.65±0.036
3G-3I	**XGBoost**	**TT**	83.46±0.368	0.848±0.006	0.69±0.012	84.63±2.12	0.69±0.042
3J-3L	**DNN**	**TT**	84.12±0.632	0.860±0.002	0.72±0.000	80.69±7.34	0.62±0.157

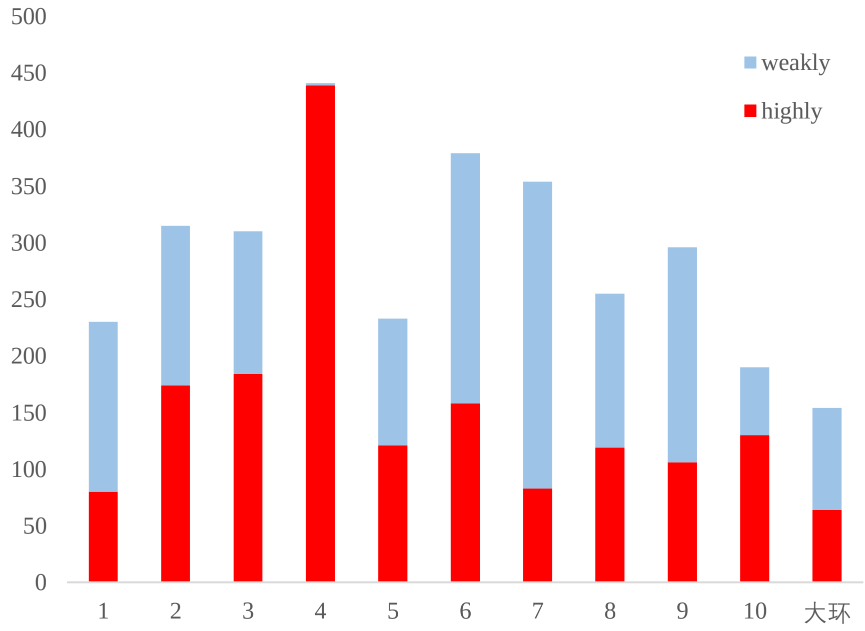
Zhao, Y., et al. Yan, A.* Classification of FLT3 inhibitors and SAR analysis by machine learning methods. *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10640-8

Classification of FLT3 inhibitors and SAR analysis by machine learning methods

Combined with the dendrogram generated by the DT algorithm and clustering via the K-means algorithm, we found some substructures were significantly related to inhibition activity. These structural fragments, core scaffolds, and side chains greatly affect the activity of the compounds against FLT3.



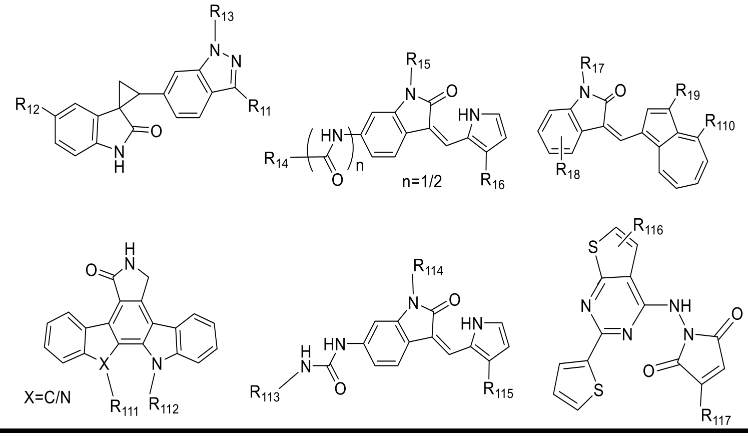
The 3264 FLT3 inhibitors (except for macrocyclic compounds) were clustered into 10 subsets by K-Means. T-SNE1 and T-SNE2 are the two dimensions reduced from the 1024 ECFP4 fingerprints by T-SNE. The red ‘X’ markers represent the compounds closest to the cluster center in each subset.



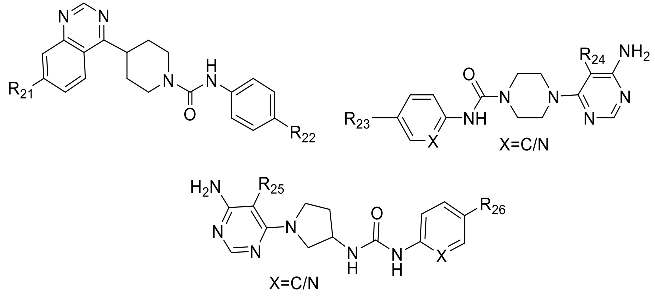
Zhao, Y., et al. Yan, A.* Classification of FLT3 inhibitors and SAR analysis by machine learning methods. *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10640-8

Classification of FLT3 inhibitors and SAR analysis by machine learning methods

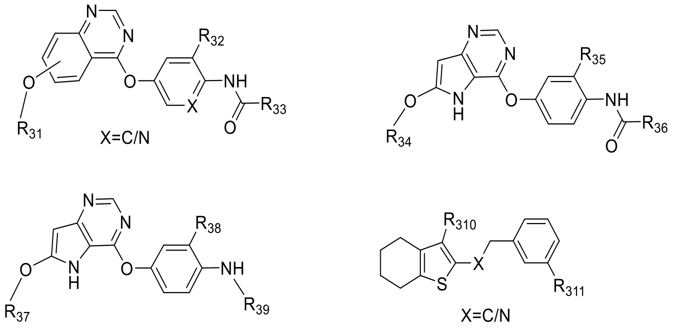
Subset 1
(80/150)



Subset 2
(174/241)

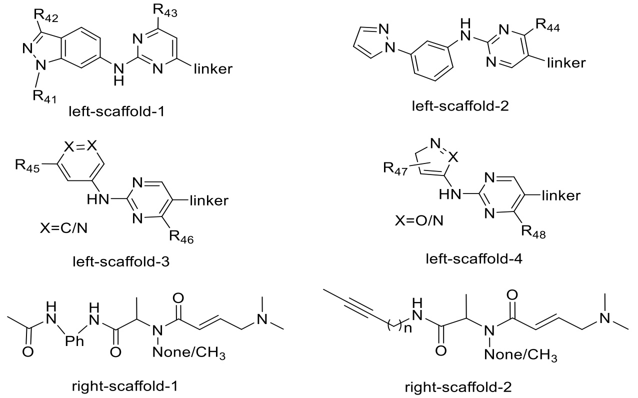
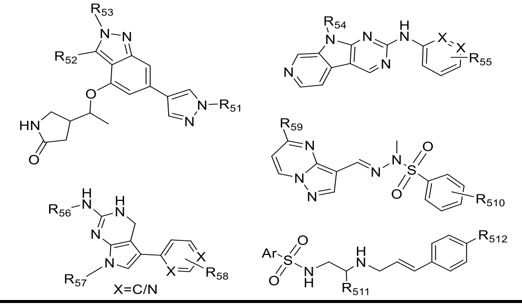
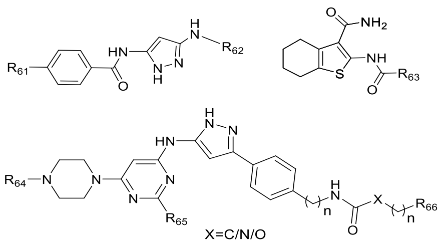


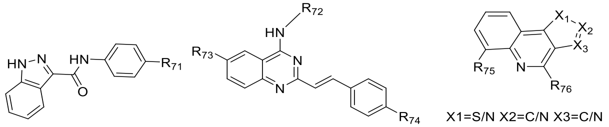
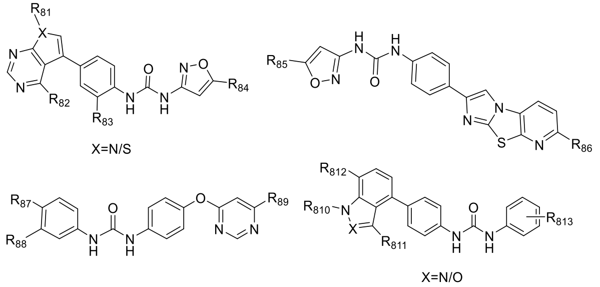
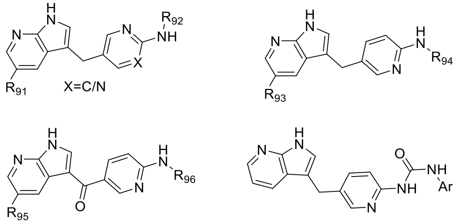
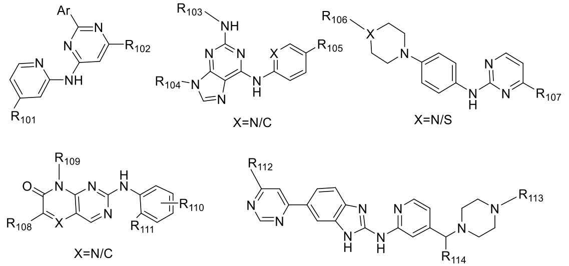
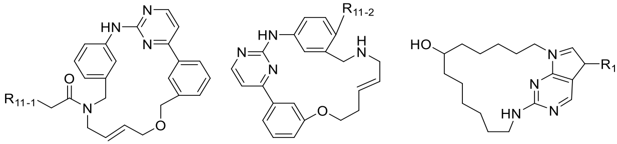
Subset 3
(184/226)



Zhao, Y., et al. Yan, A.* Classification of FLT3 inhibitors and SAR analysis by machine learning methods. *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10640-8

Classification of FLT3 inhibitors and SAR analysis by machine learning methods

Subset 4 (439/2)	 left-scaffold-1 left-scaffold-2 X=C/N left-scaffold-3 X=O/N left-scaffold-4 None/CH3 right-scaffold-1 right-scaffold-2
Subset 5 (121/112)	
Subset 6 (158/221)	

Subset 7 (146/271)	
Subset 8 (177/136)	
Subset 9 (130/106)	
Subset 10 (130/60)	
Subset 11 (64/90)	

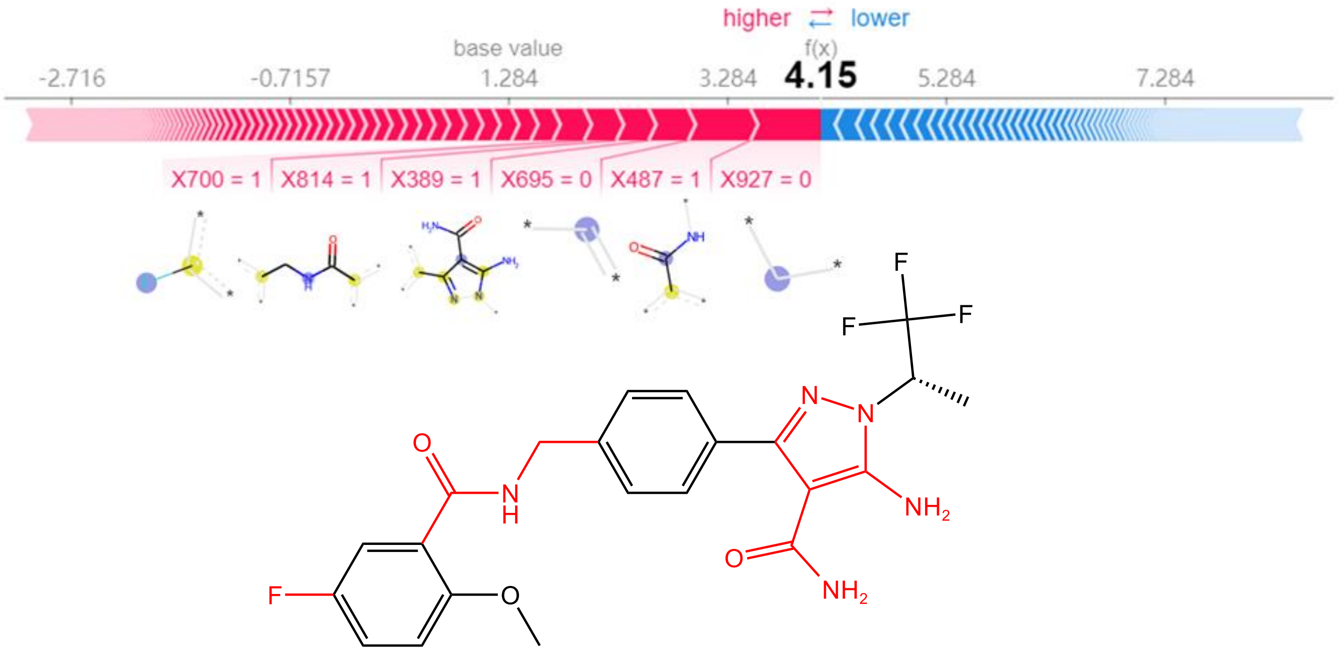
Zhao, Y., et al. Yan, A.* Classification of FLT3 inhibitors and SAR analysis by machine learning methods. *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10640-8

Prediction of bioactivities of small molecules using machine learning methods

- Classification of FLT3 inhibitors and SAR analysis by machine learning methods
- Classification of non-covalent Bruton's tyrosine kinase (BTK) inhibitors and SAR analysis by machine learning methods
- A SAR and QSAR study on cyclin dependent kinase 4 (CDK4) inhibitors using machine learning methods

Classification of non-covalent BTK inhibitors and SAR analysis by machine learning methods

We aimed to develop a predictive model capable of classifying highly and weakly active non-covalent BTK inhibitors (3895 compounds). To achieve this, we employed a suite of machine learning algorithms, including Decision Trees (DT), Random Forests (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost). Furthermore, to validate and interpret our model, we incorporated the SHAP (SHapley Additive exPlanations) method, which provided insightful explanations for our predictive outcomes.

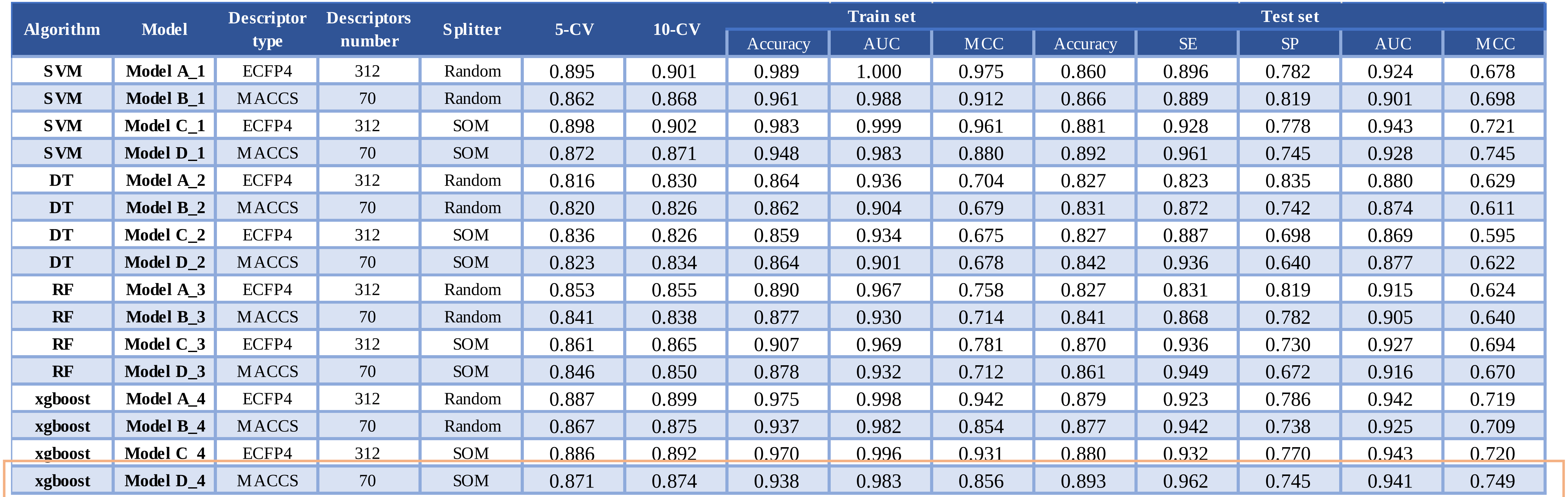


LOXO-305（Phase III，Carna Biosciences）
IC50 = 10.3 nM
ModelC_4（Xgboost）

Li, G.; et al. Yan, A.* Machine learning based classification models for non covalent Bruton's tyrosine kinase inhibitors: predictive ability and interpretability, *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10696-6

Classification of non-covalent BTK inhibitors and SAR analysis by machine learning methods

We have constructed 16 classification models for discriminating highly and weakly active non-covalent BTK inhibitors. The best model, Model D_4, which was built using XGBoost and MACCS fingerprints, achieved an accuracy of 94.1% and a Matthews correlation coefficient (MCC) of 0.75 on the test set.

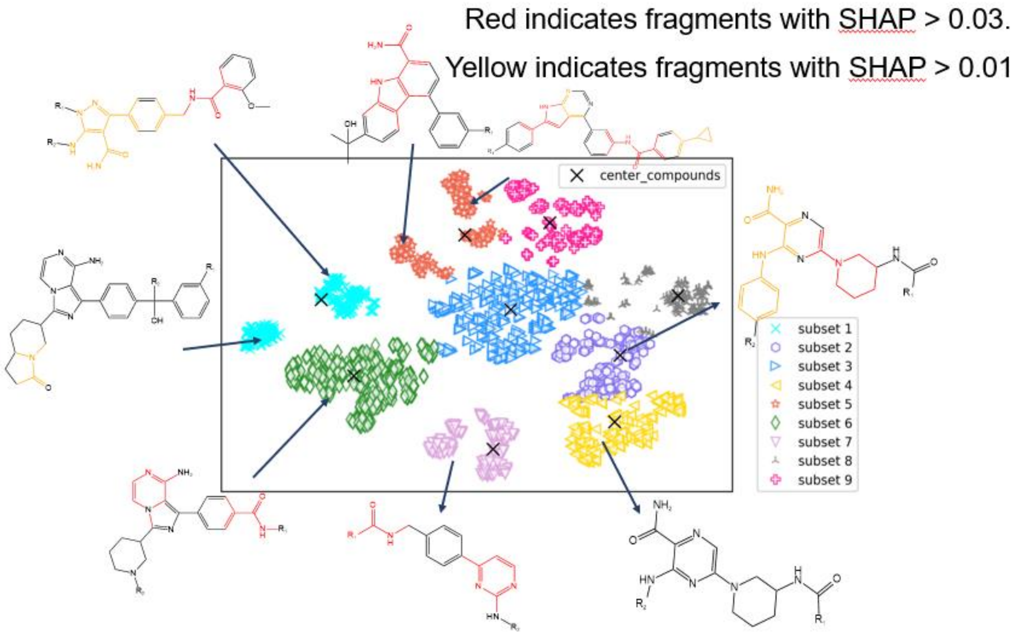


Algorithm	Model	Descriptor type	Descriptors number	Splitter	5-CV	10-CV	Train set			Test set				
							Accuracy	AUC	MCC	Accuracy	SE	SP	AUC	MCC
SVM	**Model A_1**	ECFP4	312	Random	0.895	0.901	0.989	1.000	0.975	0.860	0.896	0.782	0.924	0.678
SVM	**Model B_1**	MACCS	70	Random	0.862	0.868	0.961	0.988	0.912	0.866	0.889	0.819	0.901	0.698
SVM	**Model C_1**	ECFP4	312	SOM	0.898	0.902	0.983	0.999	0.961	0.881	0.928	0.778	0.943	0.721
SVM	**Model D_1**	MACCS	70	SOM	0.872	0.871	0.948	0.983	0.880	0.892	0.961	0.745	0.928	0.745
DT	**Model A_2**	ECFP4	312	Random	0.816	0.830	0.864	0.936	0.704	0.827	0.823	0.835	0.880	0.629
DT	**Model B_2**	MACCS	70	Random	0.820	0.826	0.862	0.904	0.679	0.831	0.872	0.742	0.874	0.611
DT	**Model C_2**	ECFP4	312	SOM	0.836	0.826	0.859	0.934	0.675	0.827	0.887	0.698	0.869	0.595
DT	**Model D_2**	MACCS	70	SOM	0.823	0.834	0.864	0.901	0.678	0.842	0.936	0.640	0.877	0.622
RF	**Model A_3**	ECFP4	312	Random	0.853	0.855	0.890	0.967	0.758	0.827	0.831	0.819	0.915	0.624
RF	**Model B_3**	MACCS	70	Random	0.841	0.838	0.877	0.930	0.714	0.841	0.868	0.782	0.905	0.640
RF	**Model C_3**	ECFP4	312	SOM	0.861	0.865	0.907	0.969	0.781	0.870	0.936	0.730	0.927	0.694
RF	**Model D_3**	MACCS	70	SOM	0.846	0.850	0.878	0.932	0.712	0.861	0.949	0.672	0.916	0.670
xgboost	**Model A_4**	ECFP4	312	Random	0.887	0.899	0.975	0.998	0.942	0.879	0.923	0.786	0.942	0.719
xgboost	**Model B_4**	MACCS	70	Random	0.867	0.875	0.937	0.982	0.854	0.877	0.942	0.738	0.925	0.709
xgboost	**Model C_4**	ECFP4	312	SOM	0.886	0.892	0.970	0.996	0.931	0.880	0.932	0.770	0.943	0.720
xgboost	**Model D_4**	MACCS	70	SOM	0.871	0.874	0.938	0.983	0.856	0.893	0.962	0.745	0.941	0.749

Li, G.; et al. Yan, A.* Machine learning based classification models for non covalent Bruton’s tyrosine kinase inhibitors: predictive ability and interpretability, *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10696-6

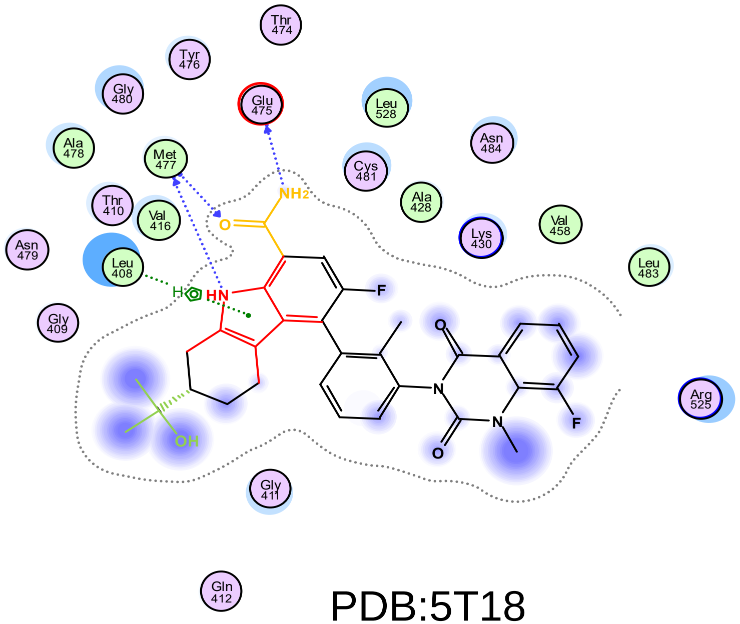
Classification of non-covalent BTK inhibitors and SAR analysis by machine learning methods

We employed the K-means clustering algorithm to identify distinct categories of non-covalent BTK inhibitors. To enhance interpretability and reliability, we used SHAP analysis to reveal the importance of different scaffold functional groups in various categories. Our approach was validated by comparing the SHAP analysis results with data from 13 existing crystal structures, confirming its effectiveness.



Red indicates fragments with SHAP > 0.03.

Yellow indicates fragments with SHAP > 0.01

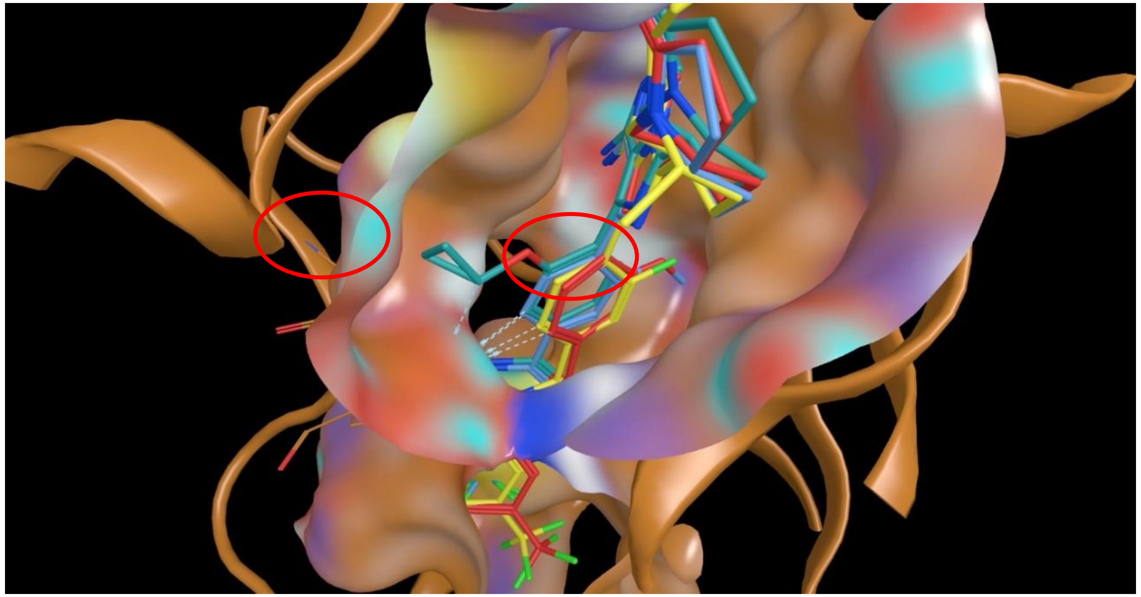


PDB:5T18

Li, G.; et al. Yan, A.* Machine learning based classification models for non covalent Bruton’s tyrosine kinase inhibitors: predictive ability and interpretability, *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10696-6

Classification of non-covalent BTK inhibitors and SAR analysis by machine learning methods

It is worth noting that in the ranking of important features of all inhibitors, X947 or X700 (F atoms at the same position) are not very important features, but for these individual ligands with crystal structures, SHAP scores very high values. Despite F atoms (X700) and methoxy groups (X947) not directly interacting with the protein, their position significantly influences ligand-protein interaction, as shown in Figures for PDB codes 5FBO, 6X3O, and 6X3P. SHAP analysis highlights their importance, demonstrating the model's robustness, generalization, and interpretability.



(a) 6X3N (Red)
(b) 5FBO (Yellow, with X700)
(c) 6X3O (blue, with X947)
(d) 6X3P (green, with X947)

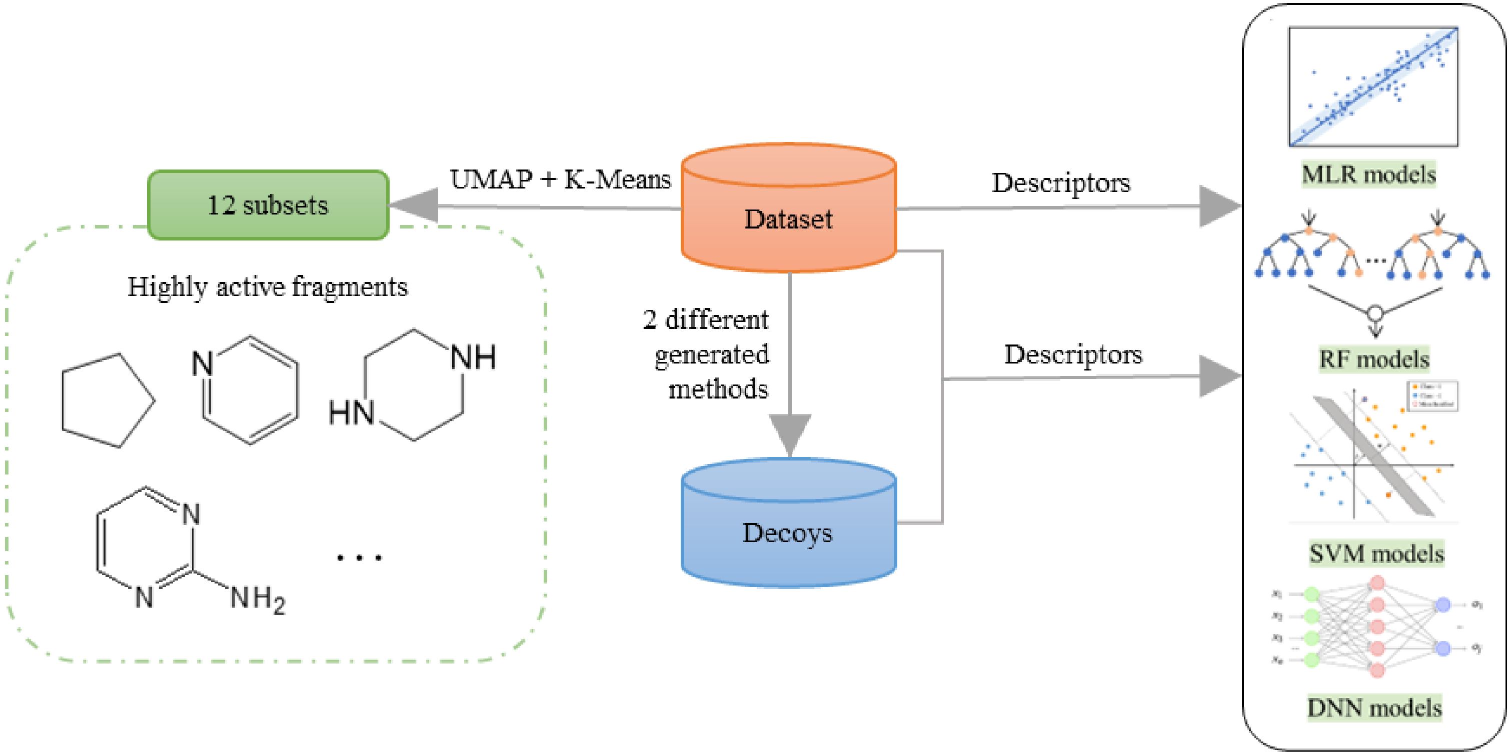
Li, G.; et al. Yan, A.* Machine learning based classification models for non covalent Bruton's tyrosine kinase inhibitors: predictive ability and interpretability, *Mol. Divers.* 2023. https://doi.org/10.1007/s11030-023-10696-6

Prediction of bioactivities of small molecules using machine learning methods

- Classification of FLT3 inhibitors and SAR analysis by machine learning methods
- Classification of non-covalent Bruton’s tyrosine kinase (BTK) inhibitors and SAR analysis by machine learning methods
- A SAR and QSAR study on cyclin dependent kinase 4 (CDK4) inhibitors using machine learning methods

A SAR and QSAR study on CDK4 inhibitors using machine learning methods

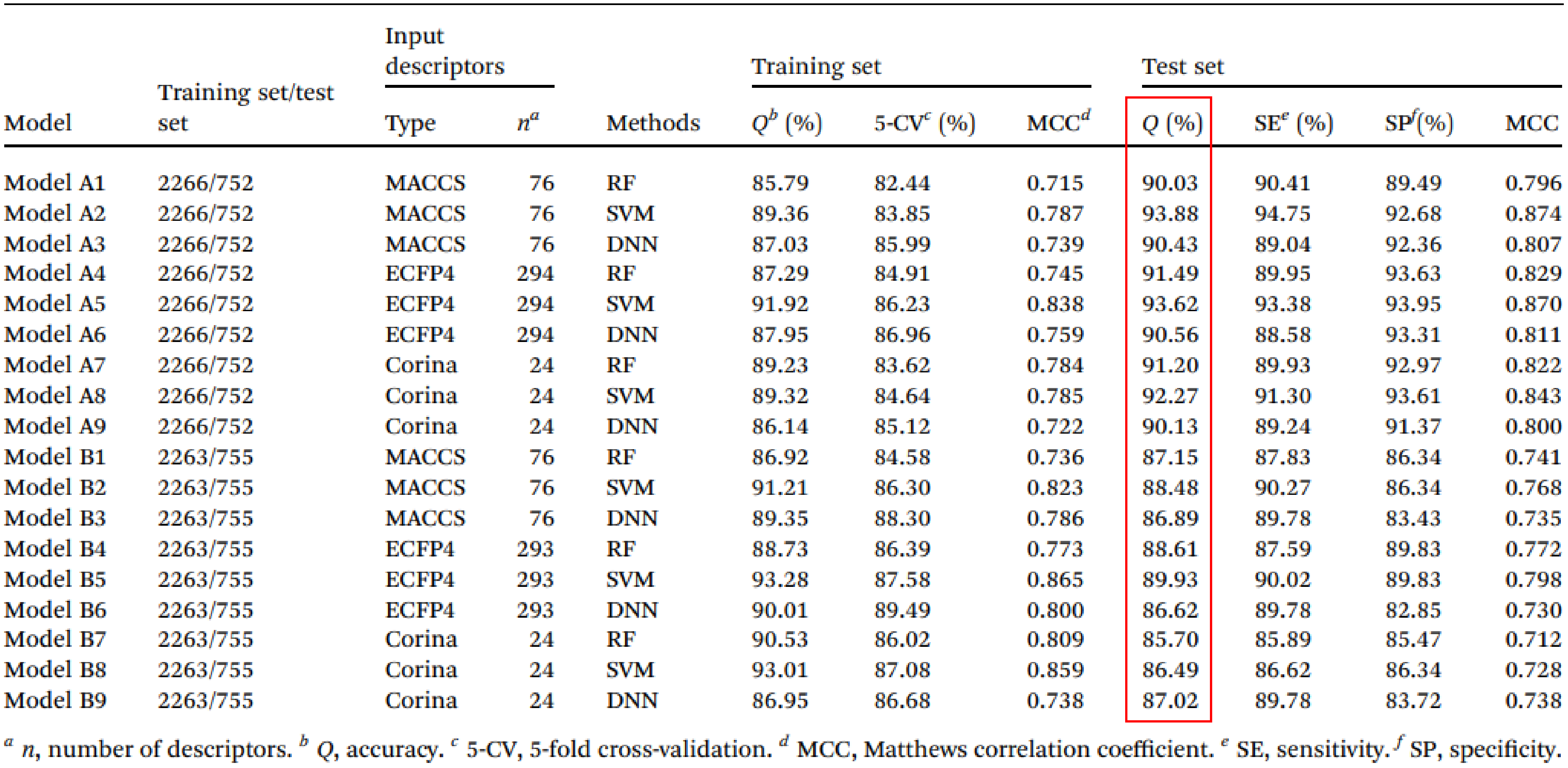
We constructed 18 classification models for discriminating highly and weakly active CDK4 inhibitors, and constructed 24 quantitative models for predicting bioactivities of CDK4 inhibitors. These models were constructed by MLR, RF, SVM and DNN algorithms, and molecules were characterized by fingerprints and molecular physicochemical descriptors. In addition, we clustered CDK4 inhibitors into 12 subsets, and analyzed their scaffolds and fragment features.



Pang, X.; et al. Yan, A.* A SAR and QSAR study on cyclin dependent kinase 4 inhibitors using machine learning methods. *Digital Discovery* 2023, 2, 1026.
https://doi.org/10.1039/d2dd00143h

A SAR and QSAR study on CDK4 inhibitors using machine learning methods

We have constructed 18 classification models for discriminating highly and weakly active CDK4 inhibitors. The accuracies of all the models were above 85%, and the highest accuracy reached 93% on the test set.



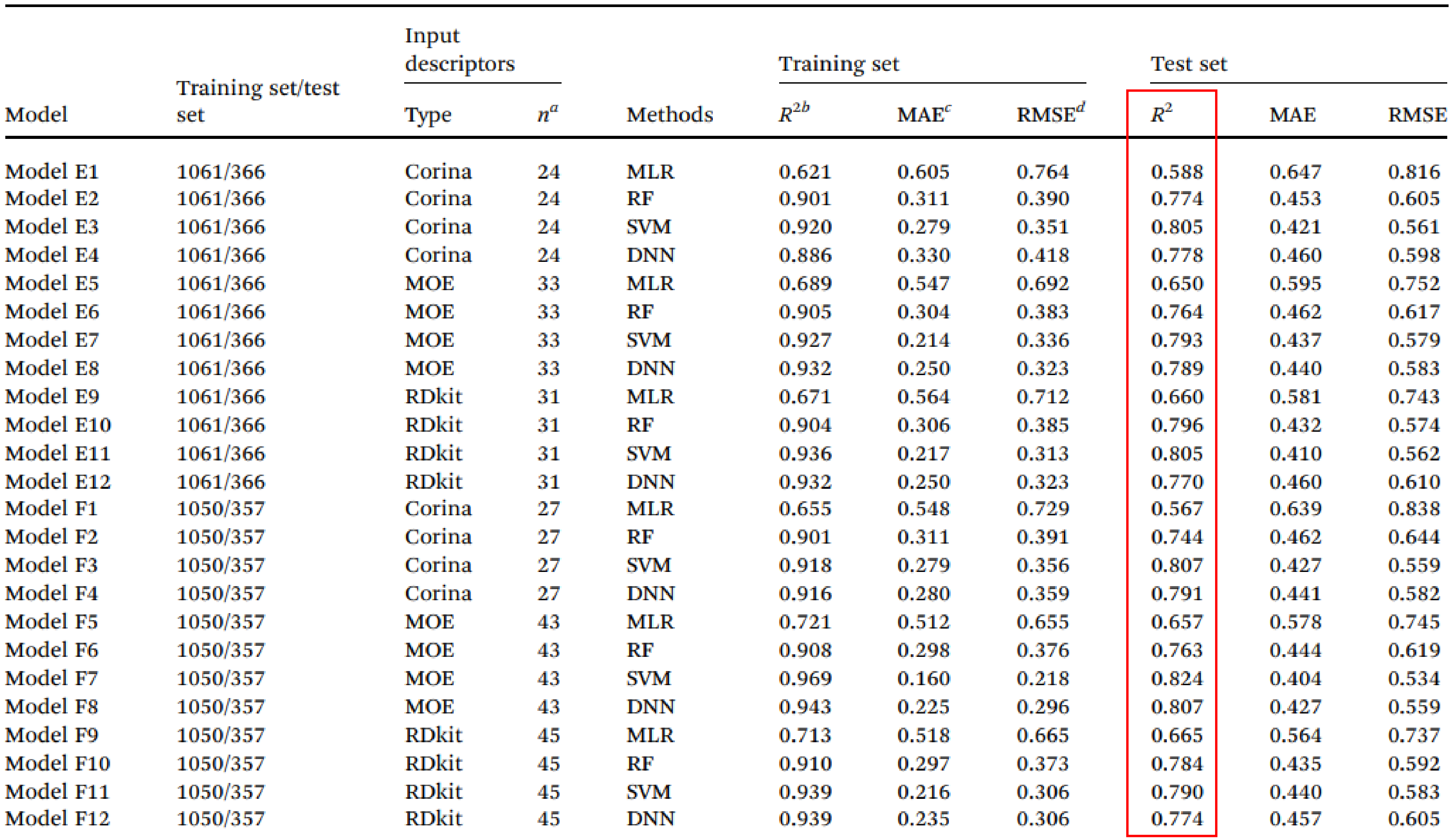
		Input descriptors			Training set			Test set			
Model	Training set/test set	Type	n^a	Methods	Q^b (%)	5-CV^c (%)	MCC^d	Q (%)	SE^e (%)	SP^f(%)	MCC
Model A1	2266/752	MACCS	76	RF	85.79	82.44	0.715	90.03	90.41	89.49	0.796
Model A2	2266/752	MACCS	76	SVM	89.36	83.85	0.787	93.88	94.75	92.68	0.874
Model A3	2266/752	MACCS	76	DNN	87.03	85.99	0.739	90.43	89.04	92.36	0.807
Model A4	2266/752	ECFP4	294	RF	87.29	84.91	0.745	91.49	89.95	93.63	0.829
Model A5	2266/752	ECFP4	294	SVM	91.92	86.23	0.838	93.62	93.38	93.95	0.870
Model A6	2266/752	ECFP4	294	DNN	87.95	86.96	0.759	90.56	88.58	93.31	0.811
Model A7	2266/752	Corina	24	RF	89.23	83.62	0.784	91.20	89.93	92.97	0.822
Model A8	2266/752	Corina	24	SVM	89.32	84.64	0.785	92.27	91.30	93.61	0.843
Model A9	2266/752	Corina	24	DNN	86.14	85.12	0.722	90.13	89.24	91.37	0.800
Model B1	2263/755	MACCS	76	RF	86.92	84.58	0.736	87.15	87.83	86.34	0.741
Model B2	2263/755	MACCS	76	SVM	91.21	86.30	0.823	88.48	90.27	86.34	0.768
Model B3	2263/755	MACCS	76	DNN	89.35	88.30	0.786	86.89	89.78	83.43	0.735
Model B4	2263/755	ECFP4	293	RF	88.73	86.39	0.773	88.61	87.59	89.83	0.772
Model B5	2263/755	ECFP4	293	SVM	93.28	87.58	0.865	89.93	90.02	89.83	0.798
Model B6	2263/755	ECFP4	293	DNN	90.01	89.49	0.800	86.62	89.78	82.85	0.730
Model B7	2263/755	Corina	24	RF	90.53	86.02	0.809	85.70	85.89	85.47	0.712
Model B8	2263/755	Corina	24	SVM	93.01	87.08	0.859	86.49	86.62	86.34	0.728
Model B9	2263/755	Corina	24	DNN	86.95	86.68	0.738	87.02	89.78	83.72	0.738

[a] n, number of descriptors. [b] Q, accuracy. [c] 5-CV, 5-fold cross-validation. [d] MCC, Matthews correlation coefficient. [e] SE, sensitivity. [f] SP, specificity.

Pang, X.; et al. Yan, A.* A SAR and QSAR study on cyclin dependent kinase 4 inhibitors using machine learning methods. *Digital Discovery* 2023, 2, 1026. https://doi.org/10.1039/d2dd00143h

A SAR and QSAR study on CDK4 inhibitors using machine learning methods

We have constructed 24 quantitative models for predicting bioactivities of CDK4 inhibitors. The R^2 values of the models based on RF, SVM and DNN algorithms were above 0.74, and the highest R^2 reached 0.82 on the test set.

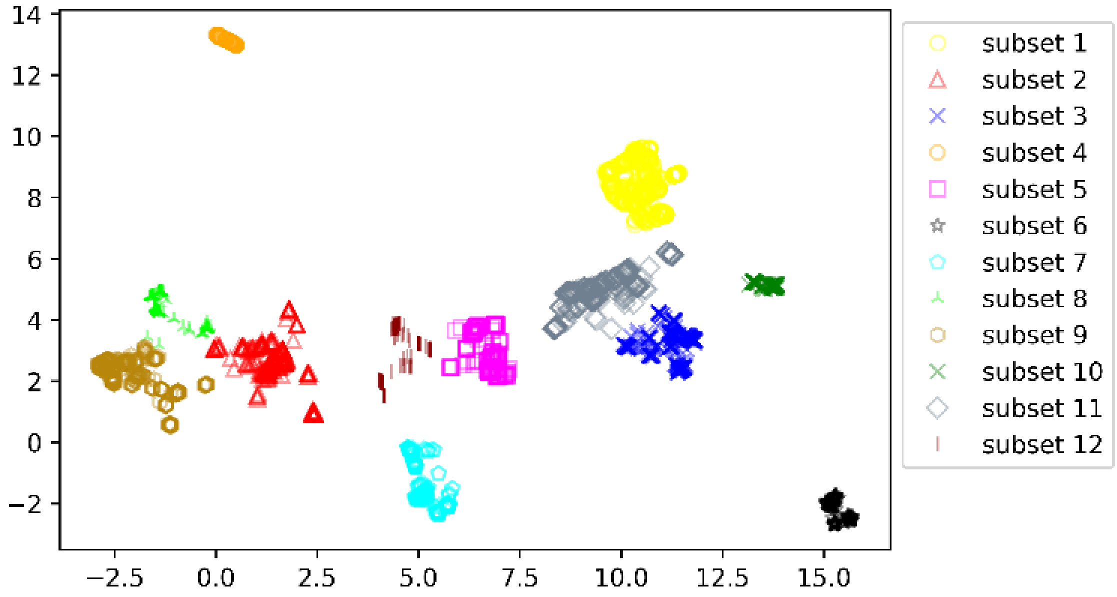


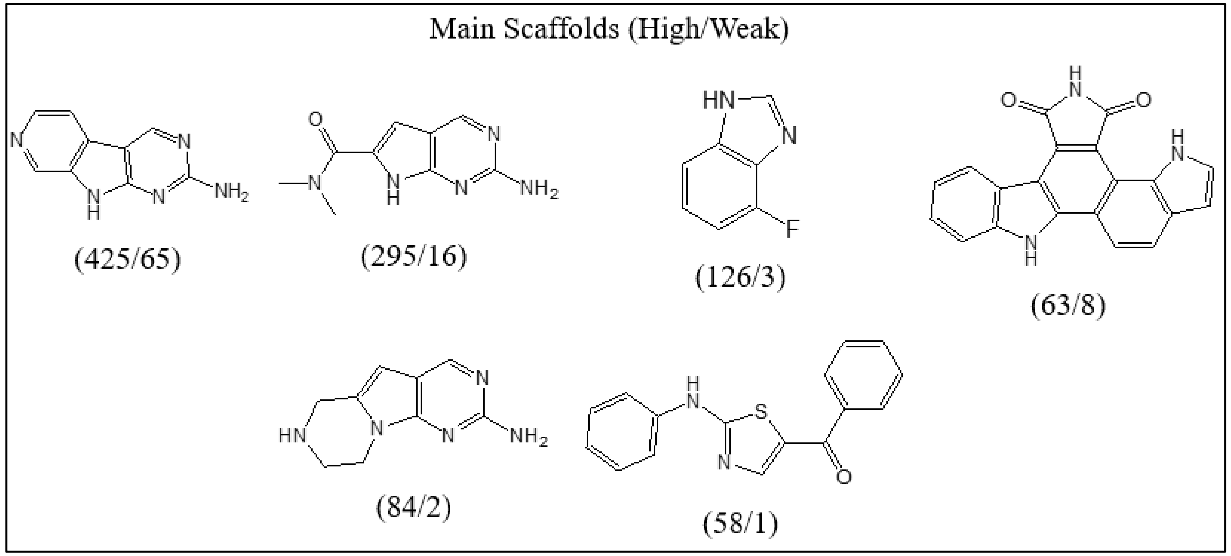
Model	Training set/test set	Input descriptors		Methods	Training set			Test set		
		Type	n^a		R^{2b}	MAE^c	RMSE^d	R^2	MAE	RMSE
Model E1	1061/366	Corina	24	MLR	0.621	0.605	0.764	0.588	0.647	0.816
Model E2	1061/366	Corina	24	RF	0.901	0.311	0.390	0.774	0.453	0.605
Model E3	1061/366	Corina	24	SVM	0.920	0.279	0.351	0.805	0.421	0.561
Model E4	1061/366	Corina	24	DNN	0.886	0.330	0.418	0.778	0.460	0.598
Model E5	1061/366	MOE	33	MLR	0.689	0.547	0.692	0.650	0.595	0.752
Model E6	1061/366	MOE	33	RF	0.905	0.304	0.383	0.764	0.462	0.617
Model E7	1061/366	MOE	33	SVM	0.927	0.214	0.336	0.793	0.437	0.579
Model E8	1061/366	MOE	33	DNN	0.932	0.250	0.323	0.789	0.440	0.583
Model E9	1061/366	RDkit	31	MLR	0.671	0.564	0.712	0.660	0.581	0.743
Model E10	1061/366	RDkit	31	RF	0.904	0.306	0.385	0.796	0.432	0.574
Model E11	1061/366	RDkit	31	SVM	0.936	0.217	0.313	0.805	0.410	0.562
Model E12	1061/366	RDkit	31	DNN	0.932	0.250	0.323	0.770	0.460	0.610
Model F1	1050/357	Corina	27	MLR	0.655	0.548	0.729	0.567	0.639	0.838
Model F2	1050/357	Corina	27	RF	0.901	0.311	0.391	0.744	0.462	0.644
Model F3	1050/357	Corina	27	SVM	0.918	0.279	0.356	0.807	0.427	0.559
Model F4	1050/357	Corina	27	DNN	0.916	0.280	0.359	0.791	0.441	0.582
Model F5	1050/357	MOE	43	MLR	0.721	0.512	0.655	0.657	0.578	0.745
Model F6	1050/357	MOE	43	RF	0.908	0.298	0.376	0.763	0.444	0.619
Model F7	1050/357	MOE	43	SVM	0.969	0.160	0.218	0.824	0.404	0.534
Model F8	1050/357	MOE	43	DNN	0.943	0.225	0.296	0.807	0.427	0.559
Model F9	1050/357	RDkit	45	MLR	0.713	0.518	0.665	0.665	0.564	0.737
Model F10	1050/357	RDkit	45	RF	0.910	0.297	0.373	0.784	0.435	0.592
Model F11	1050/357	RDkit	45	SVM	0.939	0.216	0.306	0.790	0.440	0.583
Model F12	1050/357	RDkit	45	DNN	0.939	0.235	0.306	0.774	0.457	0.605

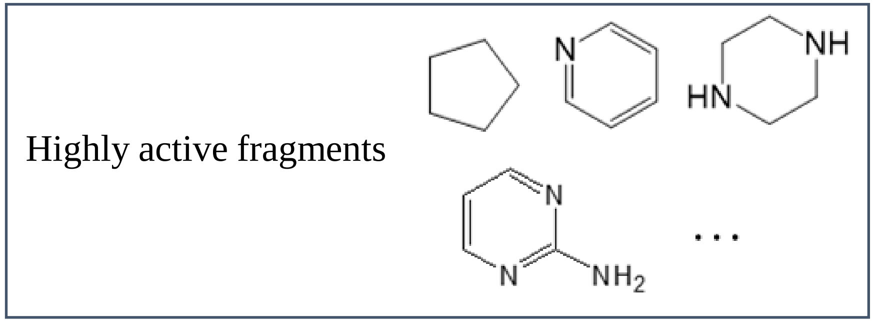
Pang, X.; et al. Yan, A.* A SAR and QSAR study on cyclin dependent kinase 4 inhibitors using machine learning methods. *Digital Discovery* 2023, 2, 1026. https://doi.org/10.1039/d2dd00143h

A SAR and QSAR study on CDK4 inhibitors using machine learning methods

In addition, we clustered CDK4 inhibitors into 12 subsets, and analyzed their scaffolds and fragment features. There were 6 scaffolds related to highly active inhibitors, and 4 important fragments in the highly active inhibitors




Main Scaffolds (High/Weak)


Highly active fragments

Pang, X.; et al. Yan, A.* A SAR and QSAR study on cyclin dependent kinase 4 inhibitors using machine learning methods. *Digital Discovery* 2023, 2, 1026. https://doi.org/10.1039/d2dd00143h

Summary

- We can use different machine learning methods (SVM, RF, DT, XGBoost, DNN) for building classification models and/or quantitative models for predicting molecular bioactivities;
- We can use different machine learning methods (K-Means, DT, SHAP) for analyzing the relationship between molecular structure and activity, and finding some key molecular fragments, substructures, and scaffolds that highly affect the bioactivity.

Acknowledgements

Miss Y. Zhao

Miss Y. Tian

Mr. J. Li

Dr. J. Liu

Miss X. Pang

Mr. G. Li

Dr. S. Shi

National Natural Science Foundation of China

“Chemical Grid Project” of Beijing University of Chemical Technology

Networks GmbH, Nuremberg, Germany