Electronic Supplementary Material (ESI) for RSC Advances.

Supplementary Information

Chemical fragments-based CDK4/6 inhibitors prediction and web server

Ling Wang,[abd] Yecheng Li,[abd] Mengyan Xu,[c] Xiaoqian Pang,[abd] Zhihong Liu,[c] Wen Tan,[*abd] and Jun Xu[*c]

[a]Guangdong Provincial Key Laboratory of Fermentation and Enzyme Engineering, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China

[b]Pre-Incubator for Innovative Drugs & Medicine, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China

[c]Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, Guangzhou 510006, China

[d]Key Laboratory of Industrial Biotechnology of Guangdong Higher Education Institutes, School of Bioscience and Bioengineering, South China University of Technology, Guangzhou 510006, China

*Correspondence to: junxu@biochemomes.com. (J. Xu); went@scut.edu.cn (W. Tan)

Contents

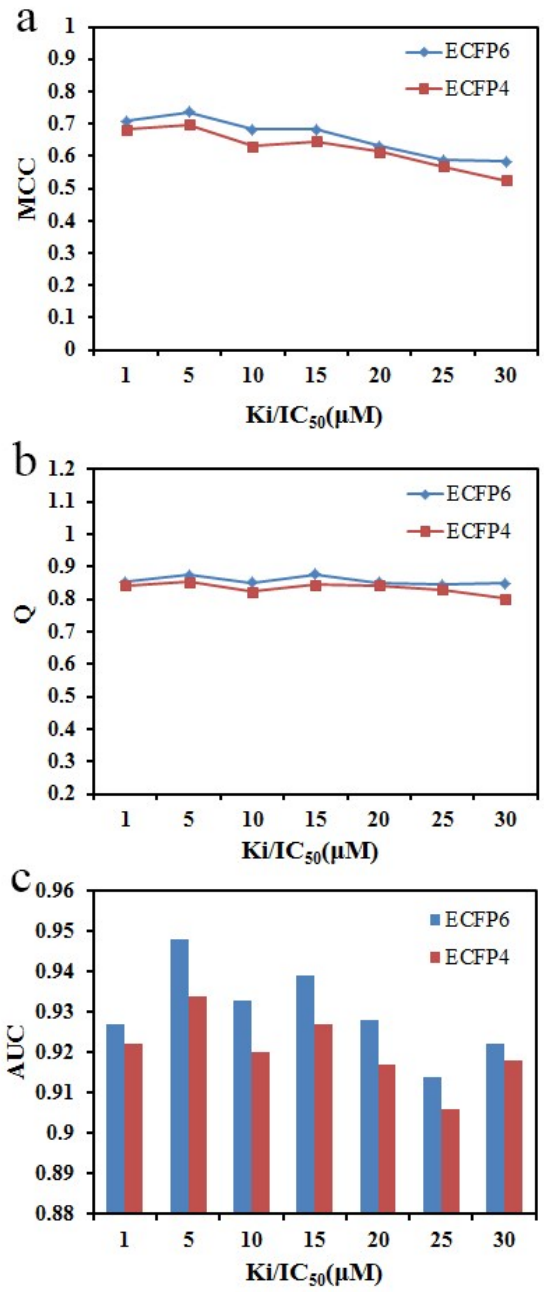

Fig. S1. The distribution of MCC (a), Q (b), and AUC values (c) based on different active cutoff values using ECFP_4 and ECFP_6 fingerprints. Q: the overall predictive accuracy; MCC: Matthews correlation coefficient; AUC: the area under the receiver operating characteristic curve.

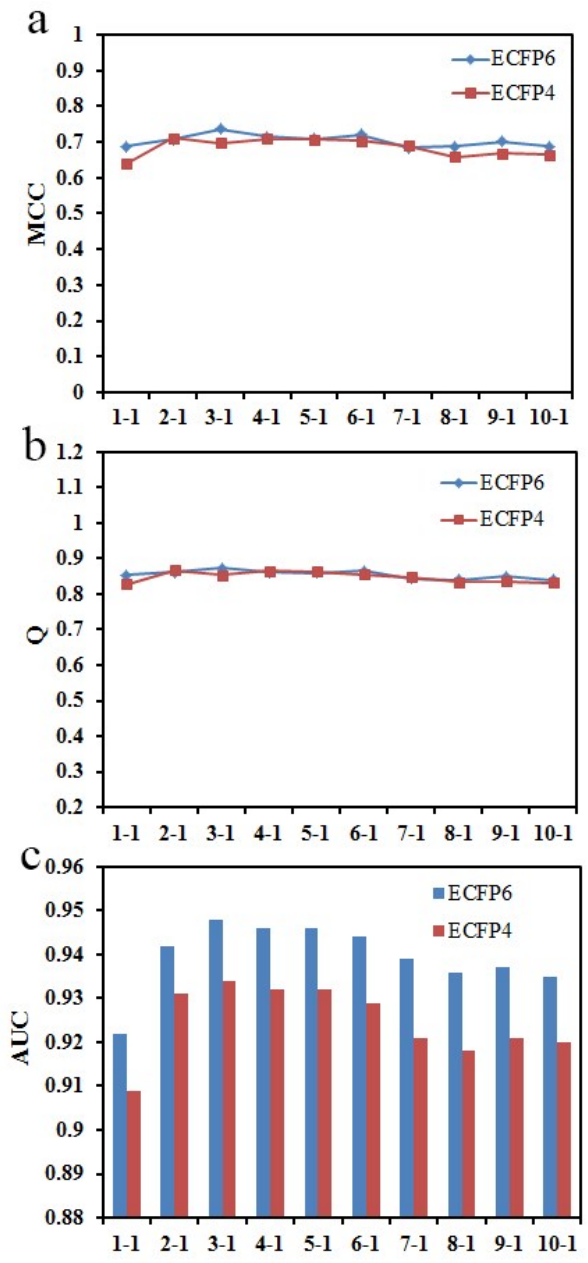

Fig. S2. The distribution of MCC (a), Q (b), and AUC values (c) based on the different proportion of the training set and test set using ECFP_4 and ECFP_6 fingerprints. Q: the overall predictive accuracy; MCC: Matthews correlation coefficient; AUC: the area under the receiver operating characteristic curve.

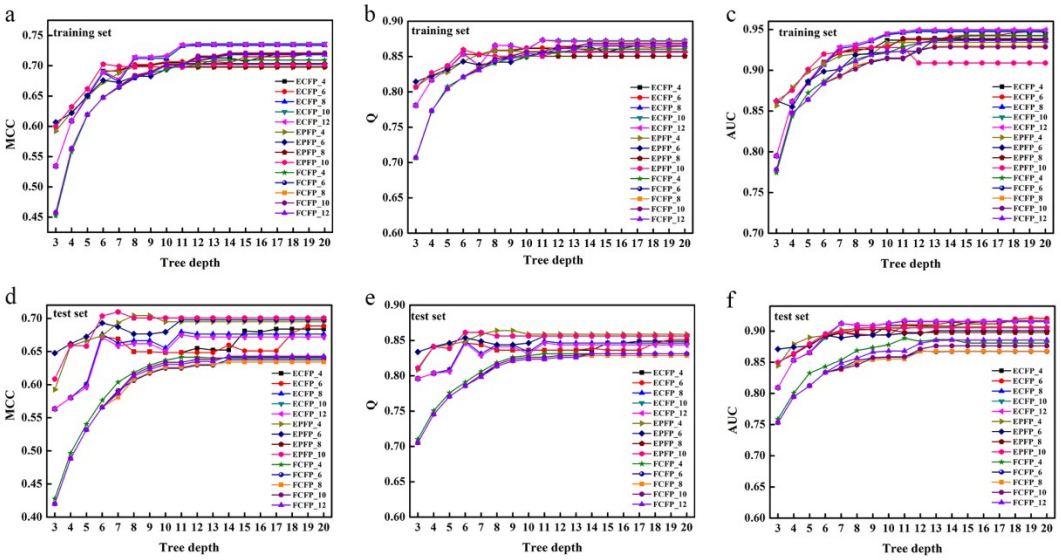

Fig. S3. The MCC, Q, and AUC of single tree models versus the tree depth of the fingerprint set (ECFP, EPFP, and FCFP) for (a,b,c) training set and (d,e,f) test set. Q: the overall predictive accuracy; MCC: Matthews correlation coefficient; AUC: the area under the receiver operating characteristic curve.

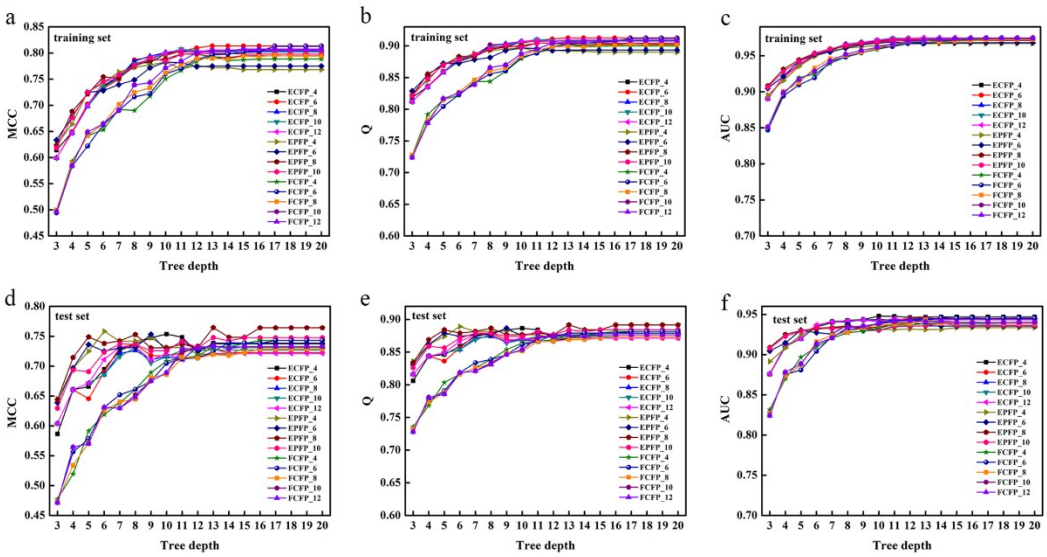

Fig. S4. The MCC, Q, and AUC of random forest models versus the tree depth of the fingerprint set (ECFP, EPFP, and FCFP) for (a,b,c) training set and (d,e,f) test set. Q: the overall predictive accuracy; MCC: Matthews correlation coefficient; AUC: the area under the receiver operating characteristic curve.

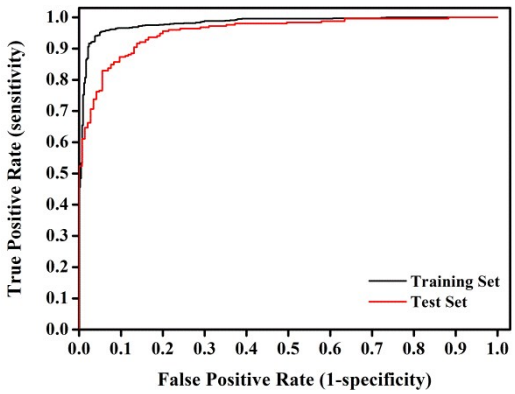

Fig. S5. The receiver operating characteristic (ROC) plot of the best Bayesian model based on LCFP_10 fingerprint for training and testing sets.

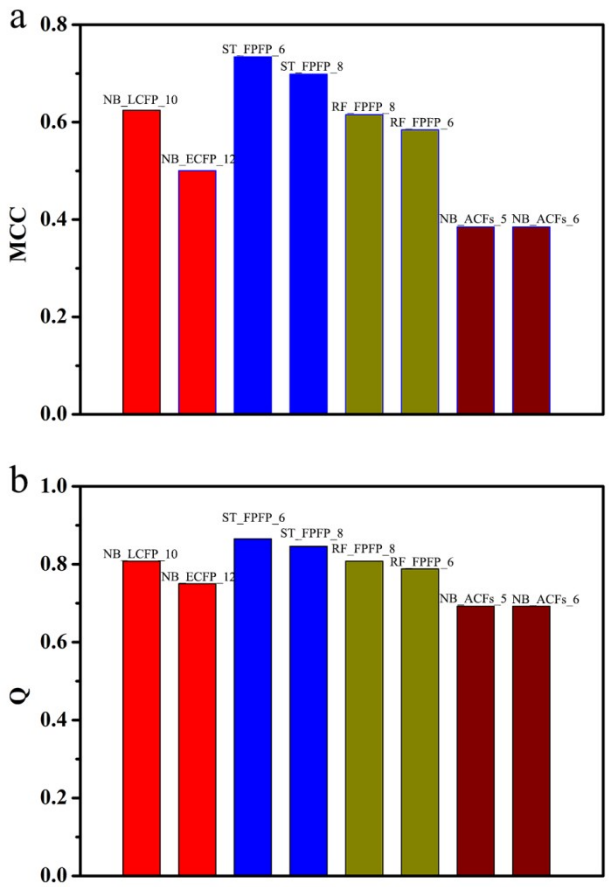

Fig. S6. The predictions for the 52 CDK6 assay data using the top two ST, RF, NB, and ACFs-NB models. MCC: Matthews correlation coefficient; Q: the overall predictive accuracy. The tree depth is 8 for ST models (FPFP_6 and FPFP_8), 7 for RF model (FPFP_6), and 15 for RF model (FPFP_8), respectively.

Table S1. The structural diversity comparison of the compounds from CDK4 data set, DrugBank, and WDI databases

Data set	Compounds	Scaffolds	Diversity（Scaffolds/Compounds）
CDK4	1,588	617	38.85%
Drugbank	6,516	2,784	42.70%
WDI	70,555	24,557	34.80%

Table S2. Performance validation results of ACFs-NB models[a]

Models	Training set									Test set								
	TP	FN	TN	FP	SE	SP	MCC	Q	AUC	TP	FN	TN	FP	SE	SP	MCC	Q	AUC
ACFs-NB (1)	695	72	235	189	0.906	0.554	0.504	0.781	0.852	230	22	88	57	0.913	0.607	0.559	0.801	0.867
ACFs-NB (2)	708	59	332	92	0.923	0.783	0.72	0.873	0.942	229	23	110	35	0.909	0.759	0.681	0.854	0.935
ACFs-NB (3)	716	51	362	62	0.934	0.854	0.792	0.905	0.961	230	22	118	27	0.913	0.814	0.732	0.877	0.941
ACFs-NB (4)	728	39	380	44	0.949	0.896	0.848	0.93	0.965	235	17	120	25	0.933	0.828	0.77	0.894	0.943
ACFs-NB (5)	732	35	386	38	0.954	0.91	0.866	0.939	0.97	240	12	121	24	0.952	0.834	0.803	0.909	0.936
ACFs-NB (6)	731	36	390	34	0.953	0.92	0.872	0.941	0.975	240	12	121	24	0.952	0.834	0.803	0.909	0.936

[a]TP: true positives; TN: true negatives; FP: false positives; FN: false negatives; SE: sensitivity; SP: specificity; Q: the overall predictive accuracy; MCC: Matthews correlation coefficient; AUC: the area under the receiver operating characteristic curve. The bracket represents the ACFs layer.