Supporting information for

Structural Analysis and Identification of False Positive Hits in Luciferase-based Assays

Zi-Yi Yang[1], Jie Dong[2], Zhi-Jiang Yang[1], Ai-Ping Lu[3], Ting-Jun Hou[4], Dong-Sheng Cao[1,3]

[1]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, 410003, P.R. China

[2]Central South University of Forestry and Technology, Changsha, 410004, P.R. China

[3]Institute for Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, P.R. China

[4]College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, Zhejiang, P. R. China

Corresponding authors

Dongsheng Cao

Email: oriental-cds@163.com

Tel: +86-731-89824761

Tingjun Hou

E-mail: tingjunhou@zju.edu.cn

Tel: +86-571-88208412

Table S1. The information of the three external validation sets.

Set	AID
Set 1	1006; 1269; 1891
Set 2	720522; 1379; 2265; 2530; 588847; 366886; 366888; 366890; 366892; 366893; 366894; 366895; 366897; 366898; 575763; 575764; 575765
Set 3	2228; 2229; 2515; 488838; 366887; 366889; 366891; 366896; 493175; 588451; 588498; 602357; 602358; 602364; 602365; 602474; 602475; 624030; 652016; 694150; 720562; 720835; 725939; 725940; 725941; 725942; 725943; 725944; 725945; 725946; 725947; 1053123; 1224835; 1347047

Table S2. The top 10 Murcko scaffolds and carbon skeletons appeared in the Non-inhibitor set and Inhibitor set.

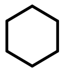
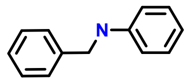
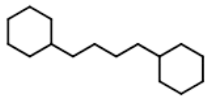
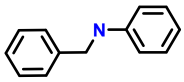
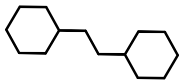
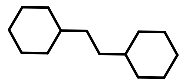
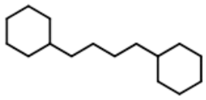
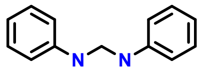
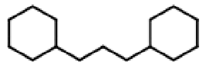
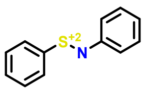
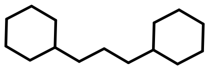
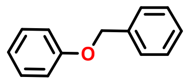
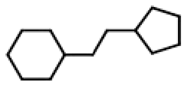
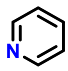
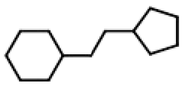
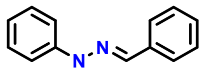
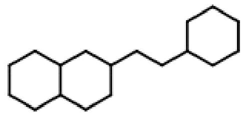
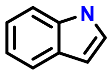
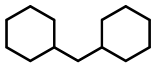
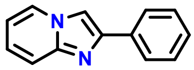
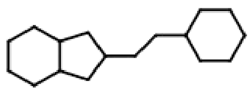
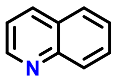
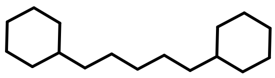
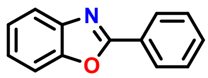
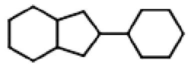
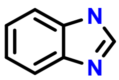
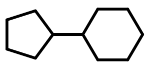
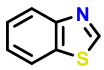
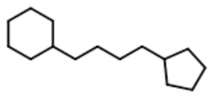
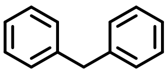
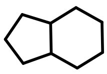
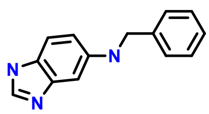
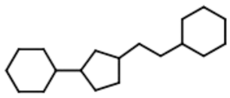
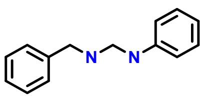
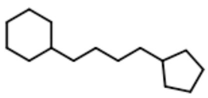
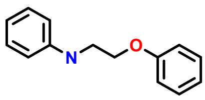
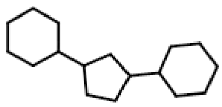
Non-inhibitor set		Inhibitor set	
Murcko Scaffold	Carbon Skeleton	Murcko Scaffold	Carbon Skeleton
			
			
Sequence only			
			
			
			
			
			
			
			

Table S3. Performance of the consensus models derived from the combinations of different ML algorithms.

		Validation set				Test set			
		BA	MCC	F1	AUC	BA	MCC	F1	AUC
ECFP4	XG+RF	0.873	0.660	0.961	0.954	0.882	0.382	0.946	0.945
	XG+DNN	0.874	0.648	0.959	0.952	0.879	0.366	0.941	0.942
	RF+DNN	0.861	0.628	0.957	0.948	0.874	0.365	0.942	0.939
	ALL	0.871	0.649	0.960	0.952	0.879	0.372	0.944	0.943
EFG	XG+RF	0.837	0.575	0.949	0.936	0.856	0.324	0.929	0.925
	XG+DNN	0.838	0.575	0.949	0.935	0.856	0.319	0.926	0.924
	RF+DNN	0.833	0.571	0.949	0.935	0.854	0.321	0.928	0.924
	ALL	0.837	0.576	0.949	0.936	0.856	0.322	0.928	0.925
MACCS	XG+RF	0.864	0.638	0.958	0.950	0.873	0.361	0.941	0.939
	XG+DNN	0.866	0.633	0.957	0.950	0.875	0.357	0.938	0.939
	RF+DNN	0.861	0.629	0.957	0.948	0.873	0.357	0.939	0.938
	ALL	0.865	0.636	0.958	0.950	0.875	0.360	0.940	0.939
MOE2d	XG+RF	0.864	0.649	0.961	0.951	0.879	0.375	0.945	0.941
	XG+DNN	0.865	0.644	0.960	0.951	0.879	0.371	0.943	0.941
	RF+DNN	0.854	0.632	0.959	0.947	0.872	0.365	0.943	0.937
	ALL	0.863	0.646	0.960	0.950	0.878	0.372	0.944	0.940

Table S4. Performance of the consensus models derived from the combinations of different types of molecular descriptors.

		Validation set				Test set			
		BA	MCC	F1	AUC	BA	MCC	F1	AUC
RF	ECFP4+EFG	0.853	0.627	0.958	0.946	0.870	0.362	0.943	0.935
	ECFP4+MACCS	0.861	0.646	0.961	0.950	0.876	0.377	0.947	0.940
	ECFP4+MOE2d	0.859	0.650	0.962	0.950	0.876	0.380	0.948	0.941
	ALL	0.856	0.637	0.960	0.948	0.872	0.367	0.944	0.937
XGBoost	ECFP4+EFG	0.869	0.648	0.960	0.951	0.876	0.363	0.941	0.940
	ECFP4+MACCS	0.877	0.661	0.961	0.956	0.885	0.377	0.944	0.945
	ECFP4+MOE2d	0.878	0.670	0.963	0.958	0.886	0.384	0.946	0.947
	ALL	0.874	0.662	0.962	0.954	0.880	0.372	0.943	0.943
DNN	ECFP4+EFG	0.859	0.626	0.956	0.944	0.867	0.345	0.936	0.934
	ECFP4+MACCS	0.867	0.634	0.957	0.949	0.876	0.358	0.939	0.939
	ECFP4+MOE2d	0.866	0.642	0.959	0.950	0.879	0.366	0.941	0.940
	ALL	0.861	0.635	0.958	0.948	0.875	0.359	0.939	0.938

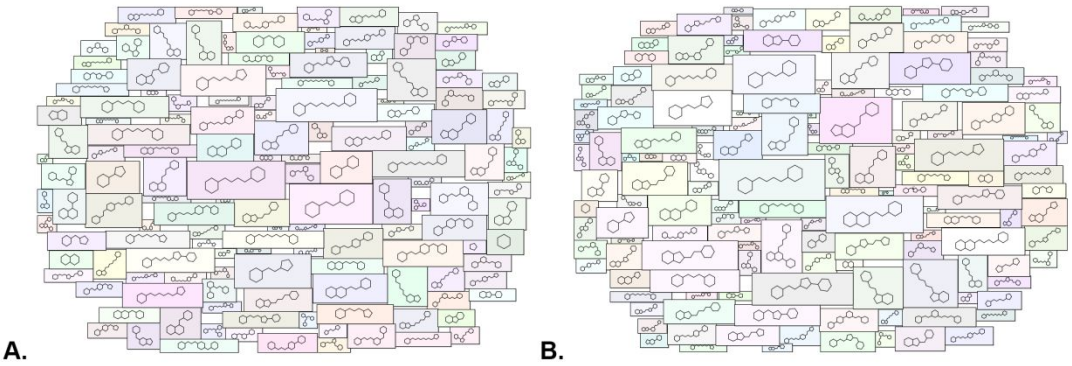


Figure S1. The cloud grams of the top 150 carbon skeletons in the (A) FLuc noninhibitor set and (B) FLuc inhibitor set.

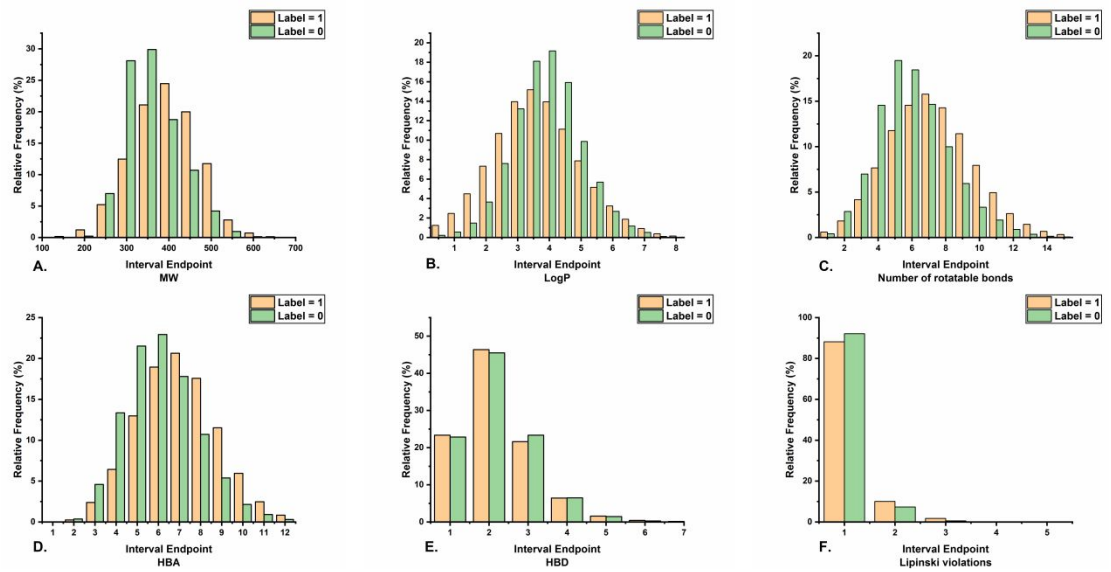


Figure S2. Distributions of (A) MW, (B) logP, (C) the number of rotatable bonds, (D) HBA, (E) HBD and (F) the violation number of Lipinski's Ro5 for the noninhibitors (label = 1) and inhibitors (label = 0) in the training set.

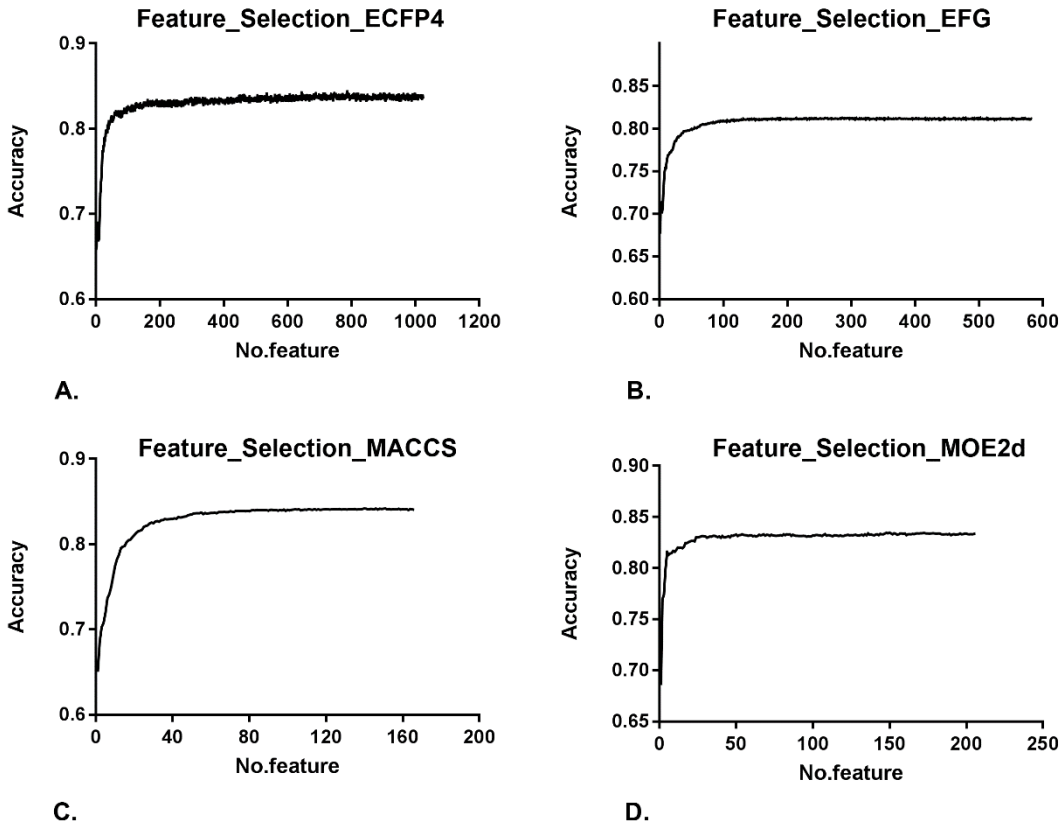


Figure S3. The feature selection results for (A). ECFP4, (B). EFG, (C). MACCS and (D) MOE2d.

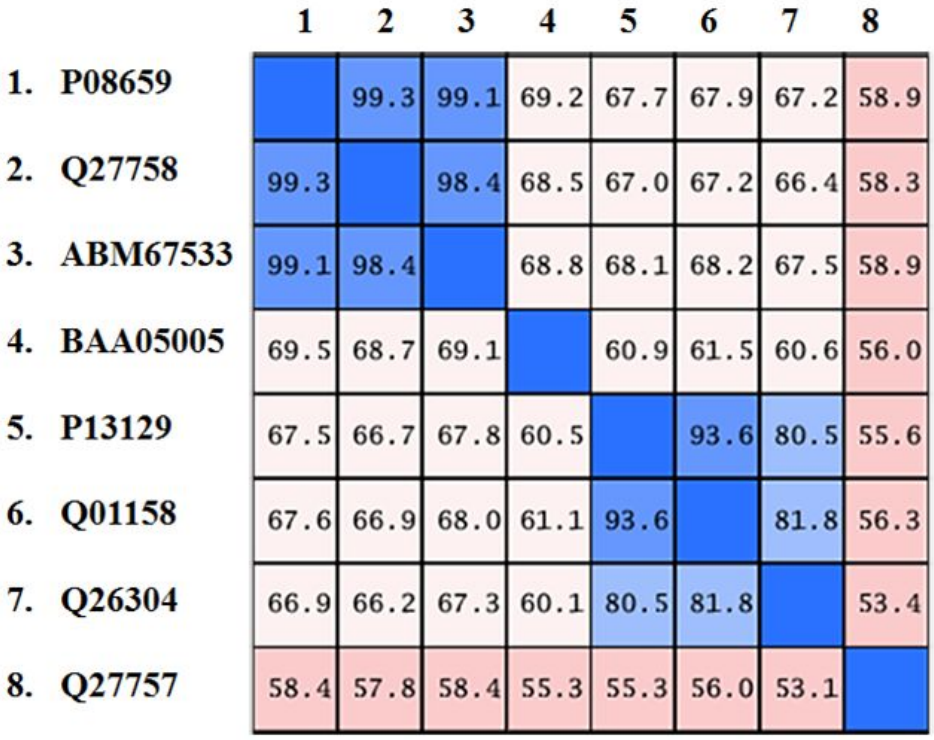


Figure S4. The sequence identity between the different subtypes of the FLuc enzymes.