Electronic Supplementary Material (ESI) for Molecular BioSystems.

In the present study, three ensemble classifiers,[1] AdaBoost, LibD3C, and Random Forest, were applied to build the classification models and validated with a 5-fold cross validation scheme.

AdaBoost (Adaptive Boosting)

AdaBoost is an ensemble method that generates a sequence of base learners focusing on the errors of previous one into a boosted classifier with weights.[2, 3] The AdaBoost M1 models were built using a software package (WEKA 3.7[4]). The JRip was selected as the base classifier and the other parameters were set as default.

LibD3C

LibD3C is a selective ensemble classifier, where multiple candidate classifiers are trained, and a set of several classifiers that are accurate and diverse are selected to deal with the problem.[5] Detailed descriptions of LibD3C can be found in literature. The LibD3C package was installed via the package manager in WEKA 3.7. The parameters were set as default.

Random Forest

Random forest is a tree-based ensemble classifier. It grows many classification trees. These trees vote to generate the most popular class.[2, 6] The random forest models were built using a software package (Orange 2.7). The number of the trees in a forest ranged from 3 to 15. The best number of the tree is the one with the highest accuracy in the testing.

The performance of the ensemble classifiers

For the property descriptor-based models, the performance of AdaBoost models was better than LibD3C and Random forest (Table S1, Table S2, and Table S3). Adaboost models were worse than descriptor-based SVM models. LibD3C and random forest models were slightly better than KNN models.

For the structural fingerprint-based models, the ES-based LibD3C model had the best predictivity in ensemble classifiers. Overall, the fingerprint-based LibD3C models were better than AdaBoost and random forest models. Fingerprint-based LibD3C models were better than descriptor-based LibD3C models. LibD3C models were worse than fingerprint-based SVM models (except SVM_ES model). Fingerprint-based LibD3C models were better than fingerprint-based KNN, RP, and NB models (except RP_S and NB_S models).

Overall, the combinatorial AdaBoost and LibD3C models were better than combinatorial random forest models. The combinatorial AdaBoost models (PaDEL_S and PaDEL_SC) and LibD3C models (MOE_MA, MOE_S, PaDEL_MA, and PaDEL_S) achieved the best overall predictivity (MCC values were greater than 0.85). ES Fingerprint-descriptor based AdaBoost and LibD3C models were worse than property descriptor based or fingerprint-based models. Overall, the predictivity of combined AdaBoost and LibD3C models was comparable to the combined SVM models.

Table S1 The performance of AdaBoost models based on property descriptors and structural fingerprints

AdaBoost Models	**Training set**								**Test set**							
	TP	**TN**	**FP**	**FN**	**SE**	**SP**	**Q**	**MCC**	**TP**	**TN**	**FP**	**FN**	**SE**	**SP**	**Q**	**MCC**
MOE	56	95	12	13	0.812	0.888	0.858	0.701	18	35	1	4	0.818	0.972	0.914	0.817
PaDEL	53	95	12	16	0.768	0.888	0.841	0.664	20	33	3	2	0.909	0.917	0.914	0.819
ES	58	98	9	11	0.841	0.916	0.886	0.761	16	33	3	6	0.727	0.917	0.845	0.666
MA	61	97	10	8	0.884	0.907	0.898	0.787	16	34	2	6	0.727	0.944	0.862	0.705
S	60	95	12	9	0.870	0.888	0.881	0.752	18	34	2	4	0.818	0.944	0.897	0.779
SC	57	98	9	12	0.826	0.916	0.881	0.748	18	34	2	4	0.818	0.944	0.897	0.779
MOE-ES	57	94	13	12	0.826	0.879	0.858	0.703	19	33	3	3	0.864	0.917	0.897	0.780
MOE-MA	55	99	8	14	0.797	0.925	0.875	0.736	19	33	3	3	0.864	0.917	0.897	0.780
MOE-S	58	91	16	11	0.841	0.850	0.847	0.683	18	35	1	4	0.818	0.972	0.914	0.817
MOE-SC	56	94	13	13	0.812	0.879	0.852	0.690	19	34	2	3	0.864	0.944	0.914	0.816
PaDEL-ES	57	96	11	12	0.826	0.897	0.869	0.725	17	34	2	5	0.773	0.944	0.879	0.741
PaDEL-MA	56	96	11	13	0.812	0.897	0.864	0.713	18	35	1	4	0.818	0.972	0.914	0.817
PaDEL -S	57	96	11	12	0.826	0.897	0.869	0.725	19	35	1	3	0.864	0.972	0.931	0.853
PaDEL -SC	56	97	10	13	0.812	0.907	0.869	0.724	21	34	2	1	0.955	0.944	0.948	0.892

Table S2 The performance of LibD3C models based on property descriptors and structural fingerprints

LibD3C Models	**Training set**								**Test set**							
	TP	**TN**	**FP**	**FN**	**SE**	**SP**	**Q**	**MCC**	**TP**	**TN**	**FP**	**FN**	**SE**	**SP**	**Q**	**MCC**
MOE	52	99	8	17	0.754	0.925	0.858	0.699	17	32	4	5	0.773	0.889	0.845	0.668
PaDEL	53	101	6	16	0.768	0.944	0.875	0.736	15	35	1	7	0.682	0.972	0.862	0.710

ES	46	101	6	23	0.667	0.944	0.835	0.653	17	36	0	5	0.773	1.000	0.914	0.824
MA	52	95	12	17	0.754	0.888	0.835	0.651	17	35	1	5	0.773	0.972	0.897	0.781
S	50	99	8	19	0.725	0.925	0.847	0.675	15	35	1	7	0.682	0.972	0.862	0.710
SC	57	92	15	12	0.826	0.860	0.847	0.681	16	36	0	6	0.727	1.000	0.897	0.790
MOE-ES	54	95	12	15	0.783	0.888	0.847	0.676	9	36	0	13	0.409	1.000	0.776	0.548
MOE-MA	49	99	8	20	0.710	0.925	0.841	0.663	19	35	1	3	0.864	0.972	0.931	0.853
MOE-S	48	95	12	21	0.696	0.888	0.813	0.601	18	36	0	4	0.818	1.000	0.931	0.858
MOE-SC	58	93	14	11	0.841	0.869	0.858	0.705	19	34	2	3	0.864	0.944	0.914	0.816
PaDEL-ES	63	96	11	6	0.913	0.897	0.903	0.801	11	35	1	11	0.500	0.972	0.793	0.566
PaDEL-MA	59	95	12	10	0.855	0.888	0.875	0.739	19	35	1	3	0.864	0.972	0.931	0.853
PaDEL-S	53	98	9	16	0.768	0.916	0.858	0.699	19	36	0	3	0.864	1.000	0.948	0.893
PaDEL-SC	55	94	13	14	0.797	0.879	0.847	0.677	18	34	2	4	0.818	0.944	0.897	0.779

Table S3 The performance of Random forest models based on property descriptors and structural fingerprints

RF Models	Training set								Test set							
	TP	TN	FP	FN	SE	SP	Q	MCC	TP	TN	FP	FN	SE	SP	Q	MCC
MOE	56	91	16	13	0.812	0.851	0.835	0.657	19	32	4	3	0.864	0.889	0.879	0.746
PaDEL	57	96	11	12	0.826	0.897	0.869	0.725	18	31	5	4	0.818	0.861	0.845	0.674
ES	59	88	19	10	0.855	0.822	0.835	0.666	19	32	4	3	0.864	0.889	0.879	0.746
MA	57	97	10	12	0.826	0.907	0.875	0.737	18	32	4	4	0.818	0.889	0.862	0.707
S	60	95	12	9	0.870	0.888	0.881	0.752	13	32	4	9	0.591	0.889	0.776	0.512
SC	59	95	12	10	0.855	0.888	0.875	0.739	17	33	3	5	0.773	0.917	0.862	0.704
MOE-ES	59	97	10	10	0.855	0.907	0.886	0.762	18	30	6	4	0.818	0.833	0.828	0.642
MOE-MA	59	96	11	10	0.855	0.897	0.881	0.750	18	32	4	4	0.818	0.889	0.862	0.707
MOE-S	58	93	14	11	0.841	0.869	0.858	0.705	18	34	2	4	0.818	0.944	0.897	0.779
MOE-SC	59	95	12	10	0.855	0.888	0.875	0.739	18	33	3	4	0.818	0.917	0.879	0.742
PaDEL-ES	62	91	16	7	0.899	0.851	0.869	0.736	19	31	5	3	0.864	0.861	0.862	0.714
PaDEL-MA	58	97	10	11	0.841	0.907	0.881	0.749	19	31	5	3	0.864	0.861	0.862	0.714
PaDEL -S	65	94	13	4	0.942	0.879	0.903	0.806	18	31	5	4	0.818	0.861	0.845	0.674

PaDEL -SC	64	95	12	5	0.928	0.888	0.903	0.804	18	32	4	4	0.818	0.889	0.862	0.707

RF: Random forest

Validated the top ensemble classifiers with the external test

Top-13 models (with MCC values exceeding 0.8 for test set) were tested using external test data. 7 out of the 13 models had overall predictive accuracies (Q) exceeding 90%. These models exhibited predictive performance exceeding 80% for the training, test, and the external test sets.

Table S4 Top 13 models (with MCC values exceeding 0.8 for test set) validated with external test data, test data, and training data.

Classifier	Descriptors	External test set		Test set	Training set
		NCP*	Q1	Q2	Q3
LibD3C	ES	45	59.21	91.38	83.52
	MOE_MA	73	96.05	93.10	84.09
	MOE_S	69	90.79	93.10	81.25
	MOE_SC	71	93.42	91.38	85.80
	PaDEL_MA	74	97.37	93.10	87.50
	PaDEL_S	55	72.37	94.83	85.80
AdaBoost	MOE	65	85.53	91.38	85.80
	PaDEL	68	89.47	91.38	84.09
	MOE_S	66	86.84	91.38	84.66
	MOE_SC	72	94.74	91.38	85.23
	PaDEL_MA	74	97.37	91.38	86.36
	PaDEL_S	68	89.47	93.10	86.93
	PaDEL_SC	72	94.74	94.83	86.93

* NCP: Number of correct predictions; Q1~3: overall predictive accuracies.

References

1 P. Y. Yang, Y. Zhou, B. Zomaya, A. Y., *Current Bioinformatics*, 2010, 5, 296-308.
2 J. C.-W. Chan and D. Paelinckx, *Remote Sens. Environ.*, 2008, 112, 2999-3011.
3 Y. S. Freund, R. E., *Jouranl of Japanese Society for Artificial Intelligence*, 1999, 14, 771-780.
4 M. F. Hall, E. Holmes, G. Pfahringer, B. Reutemann P. Witten, L. H., *SigKdd Explorations*, 11, 10-18.

5 C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan and Q. Zou, *Neurocomputing*, 2014, 123, 424-435.
6 L. Breiman, *Machine learning*, 2001, 45, 5-32.