In silico estimation of chemical aquatic toxicity on crustacean using chemical category methods

Qianqian Cao, Lin Liu, Hongbin Yang, Yingchun Cai, Weihua Li, Guixia Liu, Philip W. Lee, Yun Tang*

Shanghai Key Laboratory of New Drug Design, School of Pharmacy, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, China

Table S1 Performance of binary classification models of all crustacean using different fingerprints and modeling methods

Model	10-fold cross validation on training set					Test set				
	AUC	CA	SP	SE	F1 Score	AUC	CA	SP	SE	F1 Score
CDK-NN	0.79	0.72	0.70	0.74	0.73	0.87	0.78	0.84	0.70	0.74
CDK-CT	0.63	0.63	0.65	0.62	0.63	0.80	0.80	0.85	0.74	0.77
CDK-KNN	0.78	0.73	0.75	0.71	0.73	0.82	0.75	0.80	0.68	0.70
CDK-NB	0.80	0.73	0.70	0.77	0.74	0.90	0.83	0.82	0.84	0.82
CDK-RF	0.82	0.74	0.74	0.74	0.75	0.90	0.79	0.82	0.74	0.76
CDK-SVM	0.82	0.75	0.75	0.75	0.75	0.89	0.80	0.84	0.76	0.77
Est-NN	0.79	0.74	0.75	0.73	0.74	0.87	0.82	0.82	0.82	0.81
Est-CT	0.72	0.71	0.75	0.66	0.70	0.78	0.77	0.77	0.77	0.75
Est-KNN	0.76	0.72	0.73	0.71	0.72	0.85	0.79	0.78	0.80	0.77
Est-NB	0.77	0.71	0.73	0.69	0.71	0.83	0.80	0.81	0.78	0.78
Est-RF	0.80	0.72	0.69	0.75	0.74	0.85	0.78	0.78	0.78	0.76
Est-SVM	0.80	0.73	0.71	0.76	0.75	0.85	0.79	0.80	0.78	0.77
Ext-NN	0.79	0.73	0.75	0.71	0.73	0.85	0.79	0.78	0.80	0.77
Ext-CT	0.64	0.64	0.67	0.62	0.64	0.73	0.73	0.80	0.65	0.69
Ext-KNN	0.78	0.73	0.73	0.74	0.74	0.83	0.75	0.77	0.72	0.72
Ext-NB	0.81	0.74	0.71	0.77	0.75	0.86	0.82	0.80	0.84	0.81
Ext-RF	0.82	0.74	0.75	0.73	0.74	0.89	0.76	0.78	0.74	0.74
Ext-SVM	0.83	0.77	0.76	0.78	0.77	0.88	0.81	0.80	0.81	0.79
Gra-NN	0.80	0.73	0.73	0.73	0.73	0.85	0.80	0.85	0.74	0.77
Gra-CT	0.69	0.7	0.73	0.66	0.69	0.75	0.76	0.85	0.66	0.72
Gra-KNN	0.79	0.72	0.73	0.71	0.72	0.83	0.76	0.84	0.68	0.72
Gra-NB	0.75	0.67	0.55	0.80	0.71	0.78	0.78	0.74	0.82	0.77
Gra-RF	0.82	0.76	0.81	0.71	0.75	0.88	0.76	0.85	0.66	0.72
Gra-SVM	0.83	0.75	0.74	0.76	0.76	0.89	0.79	0.80	0.78	0.77
Mac-ANN	0.79	0.73	0.74	0.73	0.73	0.86	0.78	0.78	0.78	0.76
Mac-CT	0.71	0.71	0.71	0.72	0.72	0.74	0.73	0.69	0.77	0.72
Mac-KNN	0.80	0.72	0.72	0.72	0.72	0.89	0.76	0.77	0.76	0.74
Mac-NB	0.74	0.66	0.62	0.69	0.67	0.82	0.75	0.70	0.80	0.74
Mac-RF	0.82	0.74	0.75	0.73	0.74	0.89	0.79	0.82	0.76	0.77
Mac-SVM	0.83	0.76	0.75	0.77	0.76	0.9	0.81	0.82	0.80	0.79
Pub-ANN	0.80	0.74	0.75	0.73	0.74	0.85	0.76	0.81	0.70	0.73
Pub-CT	0.65	0.64	0.65	0.64	0.65	0.68	0.69	0.74	0.64	0.65
Pub-KNN	0.78	0.72	0.75	0.69	0.72	0.86	0.78	0.78	0.77	0.75
Pub-NB	0.77	0.69	0.70	0.69	0.69	0.80	0.72	0.79	0.62	0.66
Pub-RF	0.82	0.75	0.80	0.70	0.74	0.88	0.77	0.85	0.68	0.72
Pub-SVM	0.81	0.75	0.74	0.77	0.76	0.86	0.76	0.79	0.73	0.73
Sub-NN	0.82	0.75	0.74	0.75	0.75	0.87	0.79	0.81	0.76	0.76
Sub-CT	0.71	0.70	0.71	0.69	0.70	0.75	0.75	0.77	0.73	0.72

Sub-KNN	0.78	0.71	0.76	0.66	0.70	0.79	0.70	0.76	0.64	0.66
Sub-NB	0.78	0.72	0.68	0.76	0.73	0.82	0.72	0.74	0.69	0.68
Sub-RF	0.80	0.73	0.65	0.80	0.75	0.87	0.77	0.73	0.82	0.76
Sub-SVM	0.80	0.75	0.70	0.81	0.77	0.87	0.78	0.77	0.80	0.77

Table S2 The parameters settings of machine learning methods for models building

	FPName	RF(trees)	kNN(k)	NN(n_mid)	SVM(c)	SVM(g)
Local models	CDK	50	13	40	2.0	0.0078125
	Est	70	13	35	2048	0.00195
	Ext	80	11	20	2.0	0.0078125
	Gra	30	13	5	2.0	0.03125
	Mac	80	9	30	0.5	0.125
	Pub	90	13	15	128	0.000122
	Sub	90	11	25	2048	0.00195
Global models	CDK	90	9	20	8.0	0.00195
	Est	40	13	15	2.0	0.125
	Ext	70	13	15	32	0.000122
	Gra	70	7	20	8.0	0.00195
	Mac	40	11	10	2.0	0.125
	Pub	90	13	25	2.0	0.3125
	Sub	80	13	5	0.5	0.125

Table S3 The AD parameters and outlier counts for test set and external validation set

Variable		Test set		External validation set	
K	**Z**	**N_{OD}**	**N_{ID}**	**N_{OD}**	**N_{ID}**
3	0.8	8	157	19	227

Figure Legends

Figure S1. Tanimoto similarity index for data sets in local and global models. A: x-axis and y-axis were represented the number of 709 compounds, respectively; B: x-axis and y-axis were represented the number of 824 compounds, respectively.

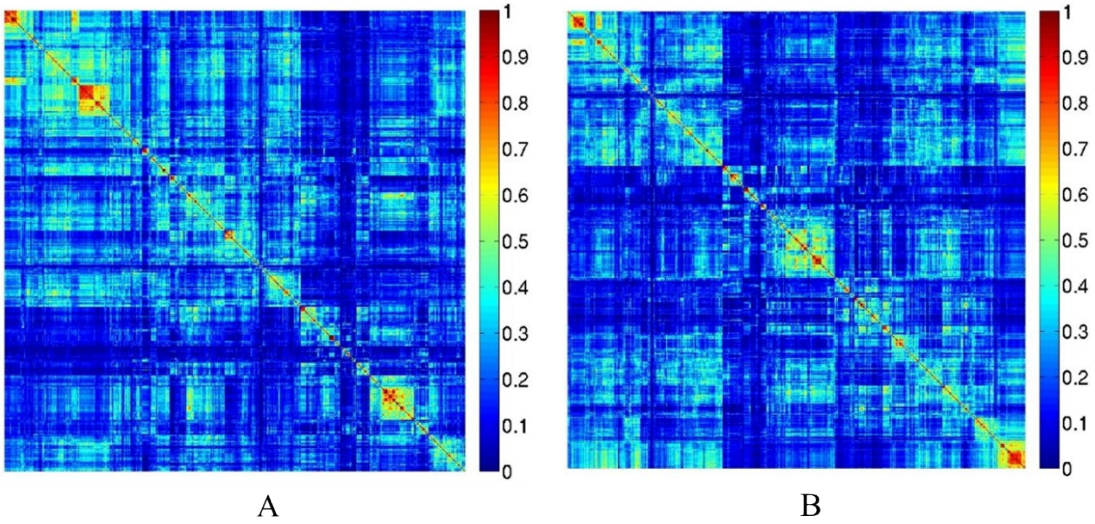

Figure S1

Figure S2. Workflow of model building for chemical acute aquatic toxicity.

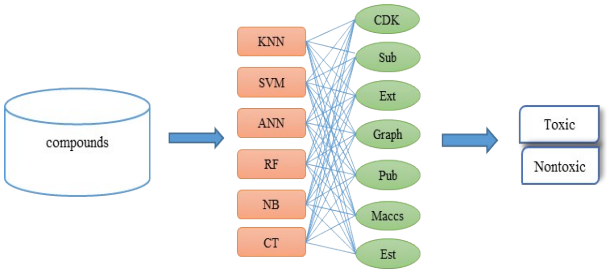

Figure S2