Supplementary Information for the Article:

admetSAR: A comprehensive source and free tool for assessment of chemical ADMET properties

Feixiong Cheng, Weihua Li, Yadi Zhou, Jie Shen, Zengrui Wu, 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.

Contact: ytang234@ecust.edu.cn

Computational Modeling Method

Molecule Description

Our recently developed substructure pattern recognition method[1] was used to depict the entire data set. Each molecule is described as a bit string structural key. The predefined dictionary contains a SMARTS list of substructure patterns. There is a one-to-one correspondence between each SMARTS pattern and each bit in the pattern fingerprint. For a SMARTS pattern, if a specified substructure is present in the given molecule, the corresponding bit is set to "1"; conversely, it is set to "0".[1] In this study, MACCS structural keys were used. The MACCS structural keys use a dictionary of MDL Public Keys[2], which contains a set of 166 most common substructure features and they are referred to as the MDL Public/MACCS keys. The definitions of MACCS structure keys are available in OpenBabel v3.11 (http://openbabel.org/).[3]

Modeling Methods

Support vector machine (SVM). Support vector machine (SVM), originally developed by Vapnik for pattern recognition, aims at minimizing the structural risk under the frame of VC theory.[4] Recently, it had been extended to the domain of regression problems.[5] In this study, support vector machine classification (SVMC) and support vector machine regression (SVMR) algorithms were selected for building classification and regression models, respectively. The classification models were built using SVM classification module provided by LIBSVM 3.11 package.[6] Regression models were built using the regression module provided by LIBSVM 2.84 package.[6, 7]

Support vector machine classification (SVMC). The classification problem can be restricted to consideration of the two-class problem without loss of generality. Detailed theory of SVM can be found in the literature.[4] Basically, in this publication, each molecule is represented using a eigenvector ***t***, and the selected patterns t_1, t_2, …, t_n make up the components of t. For SVM training, the category label y should be added. So the i^{th} molecule in the data set is defined as $M_i = (t_i, y_i)$, where $y_i = 1$ for the "positive" category and $y_i = -1$ for the "negative" category. SVM gives a decision function (classifier):

$$f(\mathbf{t}) = \mathrm{sgn}\left(\frac{1}{2}\sum\nolimits_{i=1}^{n} \alpha_i \,\mathrm{K}(\mathbf{t}_i, \mathbf{t}) + b\right). \qquad (1)$$

Where α_i is the coefficient to be learned and K is a kernel function. Parameter α_i is trained through maximizing the Lagrangian expression given below:

$$\begin{aligned} &\underset{\alpha_\mathrm{i}}{\text{maximize}} \sum\nolimits_{i=1}^{n} \alpha_\mathrm{i} - \frac{1}{2}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{n} a_i a_j y_i y_j K(\mathrm{t}_i, \mathrm{t}) \\ &\text{subject to}: \sum_{y_i=1} y_i a_i = 0, \qquad 0 \le a_i \le C; \end{aligned} \qquad (2)$$

A superiority of SVM is that it can deal with high dimensional space with the input of vectors from low dimensional space by introducing kernel function. In this study, commonly-used kernel function of Gaussian radial basis function (RBF) kernel was used. The RBF kernel has paid significant attention, most commonly with a Gaussian of the form:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \qquad (3)$$

To obtain a SVMC model with optimal performance, the penalty parameter C and different kernels parameter γ were tuned based on the training set using the grid search strategy of 5-fold cross-validation.

Probability Outputs of SVMC. Classical machine learning algorithms try to produce estimated target values (such as +1 or –1) instead of predictive probability ranges, which is easy to omit important detailed information of each classifier. In order to utilize more information of SVMC classifier, a strategy was employed to get probability output.

Lin and Weng have developed a Bayesian approach for SVM to generate probability estimation for each class in binary classification problems.[8] We briefly described how to extend probability estimation of SVMC. For probability estimation of SVMC, given *k* classes of data, for any x, the goal is to estimate:

$$p_i = p(y = i \mid x), i = 1, ..., k. \qquad (4)$$

First pairwise class probabilities are estimated:

$$r_{ij} \approx p(y = i \mid y = i \, or \, j, x) \qquad (5)$$

r_{ij} can be calculated by the following equation:

$$r_{ij} \approx \frac{1}{1 + e^{A\hat{f} + B}} \qquad (6)$$

where A and B are estimated by minimizing the negative log-likelihood function using the known training data and their decision values $\hat{f}$. Labels and decision values are required to be independent. Therefore a 5-fold cross validation was conducted to obtain the decision values. Once we have r_{ij}, we can obtain p_i by solving the following optimization problem:[8]

$$\min_{\mathrm{p}} \frac{1}{2} \sum_{i=1}^{k} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \quad \text{subject to} \quad \sum_{i=1}^{k} p_i = 1, p_i \geq 0, \forall i. \qquad (7)$$

A detailed description about solving strategy can be found in Wu’s work.[8]

Support vector machine regression (SVMR). SVM can also be applied to regression problems by the introducing an alternative loss function.[7] The loss function must be modified to include a distance measure. Using a *ε*-insensitive loss function:

$$L_\varepsilon(y) = \begin{cases} 0 & for \quad |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon & otherwise \end{cases} \tag{8}$$

In the same manner as the non-linear SVMC approach, a non-linear mapping can be used to map the data into a high dimensional feature space where linear regression is performed. The kernel approach is employed to address the curse of dimensionality. The non-linear SVMR solution, using a *ε*-insensitive loss function, which is given by:

$$\max_{a,a^*} W(a,a^*) = \max_{a,a^*} \sum_{i=1}^{l} a_i^*(y_i - \varepsilon) - a_i(y_i + \varepsilon) - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(a_i^* - a_i)(a_j^* - a_j)K(x_i, x_j) \tag{9}$$

with constraints,

$$\begin{gathered} 0 \le a_i, a_i^* \le C, \quad i = 1,\ldots,l \\ \sum_{i=1}^{l}(a_i - a_i^*) = 0. \end{gathered} \tag{10}$$

Solving Equation 9 with constraints Equation 10 determines the Lagrange multipliers, a_i, a_i^* and the regression function is given by,

$$f(x) = \sum_{\text{SVs}} (\bar{a}_\text{i} - \bar{a}_i^*)K(x_i, x) + \bar{b} \tag{11}$$

Where,

$$\begin{aligned} \langle \bar{w}, x \rangle &= \sum_{i=1}^{l}(a_i - a_i^*)K(x_i, x_j) \\ \bar{b} &= -\frac{1}{2}\sum_{i=1}^{l}(a_i - a_i^*)(K(x_i, x_r) + K(x_i, x_s)) \end{aligned} \tag{12}$$

As with the SVMR the equality constraint may be dropped if the Kernel contains *a* bias term, *b* being accommodated within the Kernel function, and the regression function is given by:

$$f(x)=\sum_{i=1}^{l}(\overline{a}_i-\overline{a}_i^*)K(x_i,x). \qquad (13)$$

A SVMR model contains three tuning parameters: Epsilon (ε) of the loss function, C of the constraints. These parameters were identified on the training set using the grid search strategy of 5-fold cross-validation.

Model Assessment Metrics

All models were validated by the 5-fold cross validation technique. The classification models were evaluated based on the counts of true positives (TP), true negatives (TN), false positives (FP), false negatives (FN). The sensitivity (*SE* = *TP*/(*TP* + *FN*)), and the specificity (SP = *TN*/(*TN* + *FP*)), were calculated. The overall accuracy (*Q*) was also calculated by the Equation 14.

$$Q=\frac{TP+TN}{TP+TN+FP+FN} \qquad (14)$$

The overall performance of regression models was evaluated by measuring the square of correlation coefficient (R^2), root mean square error (*RMSE*) calculated from the following equations:

$$R^2=1-\frac{\sum(y_i-y_j)^2}{\sum(y_i-y_m)^2} \qquad (15)$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{n_s}(y_i-y_j)^2}{n_s}} \qquad (16)$$

where, y_i, y_j and y_m represent the experimental value, predicted value and the mean of dependent variable, respectively. The n_s is the number of molecules in data set of regression equation.

In addition, a receiver operating characteristic (ROC) curve was also employed to graphically present the model behavior in a visual way. A ROC curve had been proved to be a valuable way to evaluate the quality of a binary classifier. If the area under curve (AUC) of ROC curve is 1, a perfect classifier is found, or the AUC equals 0.5, the classifier has no discriminative power at all.

The Performance of computational Models

The performance of classification models

In admetSAR, 22 highly predictive classification models, including human intestinal absorption, human oral bioavailability, blood-brain barrier penetration, P-glycoprotein substrate and inhibitor, renal organic cation transporter, volume of distribution, CPY-associated substrates and inhibition (CYP1A2, 2C9, 2C19, 2D6 and 3A4), human Ether-a-go-go-Related gene inhibition, rat acute toxicity, AMES toxicity, carcinogens, fish toxicity, Tetrahymena pyriformis toxicity, honey bee toxicity, reproductive toxicity and biodegradability, etc. were built and implemented using the SVMC algorithm. The statistics of data sets, model performance of 5-fold cross validation were given in **Table S1**.

Table S1. The statistics of data sets and detailed performance metrics of 22 classification models with probability outputs validated by 5-fold cross validation.

Model ID	Model Description	Performance Metrics			
		Q	SE	SP	AUC
A_BBB_I	The entire dataset were collected from Shen's work[1], which included 1839 compounds (1438 BBB+ and 401 BBB- compounds).	0.943	0.986	0.788	0.952
A_HIA_I	The entire dataset were collected from Shen's work[1], which included 578 compounds (500 HIA+ and 78 HIA- compounds). If a compound with the HIA% is less than 30%, it is labeled as HIA-, otherwise it is labeled as HIA+.	0.939	0.980	0.680	0.946
A_Caco2_I	In total, 674 compounds were collected, including 303 Coca2+ and 371 Coca- compounds. If a compound with the Caco-2 permeability value (Papp) ≥ 8×10^{-6} cm/s, it is labeled as high Caco-2 permeability, otherwise it is labeled as moderate-poor permeability.[9]	0.746	0.696	0.787	0.822
A_PgpS_I	In total, 332 compounds were collected from Wang's work[10], including 206 Pgp substrates and 126 Pgp non-substrates.	0.735	0.869	0.516	0.768

A_PgpI_I	In total, 1273 compounds were collected from Chen's work[11], including 797 Pgp inhibitors and 476 Pgp non-inhibitors.	0.786	0.872	0.641	0.853
A_PgpI_II	In total, 1275 compounds were collected from Broccatelli's work[12], including 666 Pgp inhibitors and 609 Pgp non-inhibitors.	0.866	0.871	0.860	0.922
M_CYP1A2I_I	In total, 14903 compounds, including 7415 inhibitors and 7488 noninhibitors were collected from Cheng's work[13]. A compound was assigned as a CYP inhibitor if the AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a noninhibitor if it has PubChem activity score equal to 0.	0.815	0.799	0.831	0.815
M_CYP2C19I_I	In total, 14576 compounds, including 6041 inhibitors and 8535 non-inhibitors were collected from Cheng's work[13]. A compound was assigned as a CYP inhibitor if the AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a non-inhibitor if it has PubChem activity score equal to 0.	0.805	0.748	0.846	0.805
M_CYP2C9I_I	In total, 14709 compounds, including 4978 inhibitors and 9731 non-inhibitors were collected from Cheng's work[13]. A compound was assigned as a CYP inhibitor if the AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a non-inhibitor if it has PubChem activity score equal to 0.	0.802	0.637	0.886	0.802
M_CYP2D6I_I	In total, 14741 compounds, including 3060 inhibitors and 11681 non-inhibitors were collected from Cheng's work[13]. A compound was assigned as a CYP inhibitor if the AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a non-inhibitor if it has PubChem activity score equal to 0.	0.855	0.456	0.960	0.855
M_CYP3A4I_I	In total, 18561 compounds, including 6707 inhibitors and 11854 non-inhibitors were collected from Cheng' work[13]. A compound was assigned as a CYP inhibitor if the AC_{50} (the compound	0.645	0.865	0.525	0.848

	concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a non-inhibitor if it has PubChem activity score equal to 0.				
M_CYPPro_I	In total, 5461 compounds, including 3269 high P450 inhibitory promiscuous compounds (I_{inh}≥0.8) and 2192 low P450 inhibitory promiscuous compounds (I_{inh}≤0.2) were collected from Cheng's work[14].	0.821	88.5	72.5	0.879
M_CYP2C9S_I	In total, 673 drugs including 142 substrates and 531 non-substrates were collected from Carbon-Mangles's work[15].	0.788	0.042	0.987	0.788
M_CYP2D6S_I	In total, 671 drugs including 191 substrates and 480 non-substrates were collected from Carbon-Mangles's work[15].	0.759	0.377	0.910	0.759
M_CYP3A4S_I	In total, 671 drugs including 357 substrates and 317 non-substrates were collected from Carbon-Mangles's work[15].	0.638	0.706	0.562	0.638
M_BIO_I	In total, 1604 diverse compounds were collected from Cheng's work.[16]	0.832	0.751	0.880	0.890
T_hERG_I	In total, 368 molecules including 79 strong hERG inhibitors (pIC50> 6.0 mol/L)and 289 weak hERG inhibitors (pIC50≤6.0 mol/L) were collected from Marchese Robinson et al[17].	0.870	0.494	0.972	0.820
T_hERG_II	In total, 806 molecules including 433 hERG inhibitors (IC50> 50 μM) and 373 hERG non-inhibitors (pIC50≤50 μM) were collected from Wang's work[18].	0.784	0.783	0.786	0.849
T_AMES_I	In total, 8445 Compounds including 4912 AMES toxic chemicals and 3533 non AMES toxic chemicals were collected from four published papers[19-21].	0.851	0.883	0.808	0.908
T_Carc_I	In total, 293 chemicals, including 64 carcinogens and 229 noncarcinogens were collected from Lagunin's work[22].	0.884	0.563	0.974	0.836
T_FHMT_I	In total, 554 compounds, including 336 high fathead minnow toxicity (FHMT) compounds and 188 low FHMT compounds were collected from EPA Fathead Minnow Acute Toxicity Database EPAFHM. If a compound with the value of LC50 more than 0.5 mmol/L were assigned as high acute FHMT compound, whereas it was assigned as low acute FHMT compounds[23].	0.814	0.896	0.654	0.880
T_HBT_I	In total, 195 pesticides or pesticide-like molecules, including 99 high honey bee toxicity (HBT) compounds and 96 low HBT compounds were collected from US EPA ECOTOX Database. If a compound with the value of LD50 more than 100μg/bee were assigned as high acute HBT compound, while it was assigned as low acute HBT compound[23].	0.759	0.758	0.760	0.824
T_TPT_I	In total, 1571 compounds, including 1217 high Tetrahymena Pyriformis Toxicity (TPT) compounds and 354 low TPT	0.917	0.958	0.776	0.956

	compounds, were collected from Cheng's work[24]. If a compound with the $pIGC_{50}$ (the negative logarithm of 50% growth inhibitory concentration) > -0.5 was assigned as TPT, otherwise as non-TPT.	

SE: Sensitivity, SP: Specificity, Q: the overall predictive accuracy, AUC: the area under the receiver operating characteristic curve.

The Performance of Regression Models

In admetSAR, five highly predictive regression models including Caco-2 permeability (LogPapp), water solubility (LogS), rat acute toxicity (LD_{50}), Tetrahymena pyriformis toxicity ($pIGC_{50}$) and Fathead Minnow Acute Toxicity (pLC_{50}) prediction were built using SVM regression algorithm and implemented. The statistics of data sets and detailed performance metrics of 5-fold cross validation were given in **Table S2**.

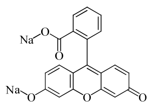
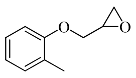
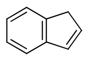
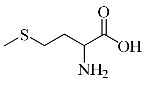
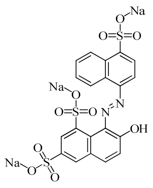
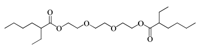
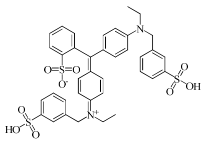
Table S2. The statistics of data sets and detailed performance of 5 regression models built using support vector regression algorithm validated by 5-fold cross validation.

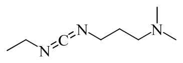
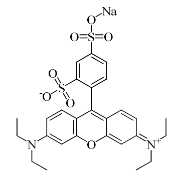
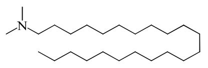
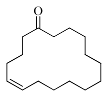
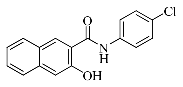
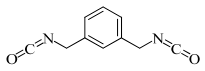
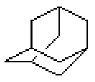
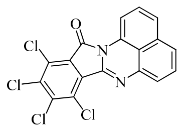
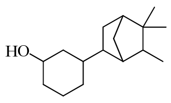
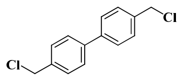
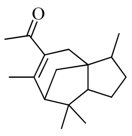
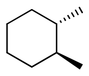
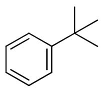
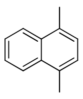
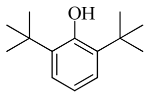
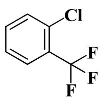
Model ID	Description	Performance Metrics		Endpoint
		RMSE	R^2	
R_A_Caco2_I	In total, 674 drug or drug-like molecules with Caco-2 permeability values were used [9].	0.339	0.564	LogPapp (cm/s)
R_A_WS_I	In total, 1708 molecules with LogS value were collected from Wang's work [25].	0.823	0.810	LogS
R_T_TPT_I	In total, 1571 compounds with $pIGC_{50}$ (ug/L) value against Tetrahymena pyriformis were collected from Cheng's work [24].	0.256	0.761	$pIGC_{50}$ (ug/L)
R_T_FHMT_I	In total, 554 pesticides or pesticide-like molecules with pLC_{50} (mg/L) value were collected from EPA Fathead Minnow Acute Toxicity Database EPAFHM [23].	0.666	0.574	pLC_{50} (mg/L)
R_T_RAT_I	In total, 10207 molecules with LD_{50} (mg/L) against rat were collected from Zhu's work [26].	0.324	0.613	LD_{50} (mol/kg)

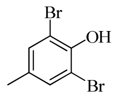
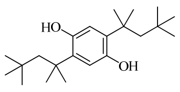
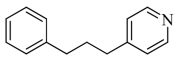
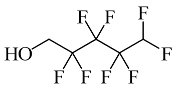
Case Study

Generalization ability of a model decides the usefulness and reliability of models. In order to test the actual predictive ability of admetSAR, the biodegradation of 27 novel chemicals was predicted firstly using admetSAR and were further assayed using the MITI-I test protocol[16]. The detailed experimental and predicted results were given in **Scheme S1** and **Table S3**. The overall predictive accuracy of admetSAR was 88.9%, that is, 24 chemicals were predicted correctly. The admetSAR outperformed *Biowin5* and *Biowin6* implmented in the EPI Suite v4.10 (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm). In addition, 9 classification models were validated using the available external validation sets (**Table S4**). And high predictive accuracies were also yielded for the external validation sets.

Scheme S1. The detailed predicted results of the admetSAR and experimental results using OECD MITI test protocol for 27 novel chemicals.

CAS RN	Structure	Indirect Analysis	Direct Analysis				Exper Results	admet SAR	*Biowin5	*Biowin6
		BOD [%]	TOC*1 [%]	UV*2 [%]	GC*2 [%]	HPLC*2 [%]				
518-47-8		0	0	-	-	0	NRB	**RB**	RB	NRB
2210-79-9		0	2	-	-	90	NRB	NRB	RB	RB
95-13-6		0	-	-	-	1	NRB	NRB	NRB	NBR
59-51-8		81	82	-	-	89	RB	RB	RB	NRB
2611-82-7		2	2	-	-	0	NRB	NRB	NRB	NRB
94-28-0		92	-	-	-	100	RB	RB	RB	RB
2650-18-2		2	0	-	-	0	NRB	NRB	NRB	NRB

1892-57-5		0	4	-	-	0	NRB	NRB	NRB	NRB
3520-42-1		6	0	-	-	0	NRB	NRB	NRB	NRB
21542-96-1		36 35	- -	- -	66 >99	- -	RB RB	NRB	RB	RB
37609-25-9		66	-	-	-	92	RB	RB	NRB	NRB
92-78-4		1	-	-	-	0	NRB	NRB	NRB	NRB
3634-83-1		0	-	-	-	>99	NRB	NRB	NRB	NRB
281-23-2		15	-	-	0	-	NRB	NRB	NRB	NRB
20749-68-2		0	-	-	-	1	NRB	NRB	NRB	NRB
3407-42-9		0	-	-	3	-	NRB	NRB	NRB	NRB
1667-10-3		0	-	-	-	3	NRB	NRB	NRB	NRB
32388-55-9		0	-	-	3	-	NRB	NRB	NRB	NRB
583-57-3		0	-	-	2	-	NRB	RB	NRB	RB
98-06-6		0	-	-	27	-	NRB	NRB	NRB	NRB
571-58-4		0	-	-	-	2	NRB	NRB	NRB	NRB
128-39-2		0	-	-	-	11	NRB	NRB	NRB	NRB
88-16-4		0	-	-	0	-	NRB	NRB	NRB	NRB

2432-14-6		0	-	-	-	0	NRB	NRB	NRB	NRB
903-19-5		0	-	-	1	-	NRB	NRB	NRB	NRB
2057-49-0		4	-	-	23	-	NRB	NRB	NRB	NRB
355-80-6		0	-	-	0	-	NRB	NRB	RB	NRB

RB: ready biodegradability, NRB: not ready biodegradability, BOD: biological oxygen demand. [*]*Biowin5* and *Biowin6* are the Linear and Non-linear MITI Biodegradation Models respectively published by Tunkel et al.,[27] which had be implmented in the EPI Suite v4.10 (http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm).

Table S3. The performance of admetSAR when predicting the biodegradability of 27 novel compounds.

Model	TP	TN	FP	FN	SE (%)	SP (%)	Q (%)
admetSAR	3	23	2	1	75.0	91.3	88.9
**Biowin5*	3	20	3	1	75.0	87.0	85.2
**Biowin6*	2	21	2	2	50.0	91.3	85.2

TP: true positives, TN: true negatives, FP: false positives, FN: false negatives, SE: Sensitivity, SP: Specificity, Q: the overall predictive accuracy,

**Biowin5* and *Biowin6* are the Linear and Non-linear MITI Biodegradation Models respectively published by Tunkel et al.,[27] which had be implmented in the EPI Suite v4.10 (http://www.epa.gov/oppt/exposure/pubs/ episuitedl.htm).

Table S4. The statistics of data sets and detailed performance metrics of 9 classification models with probability outputs validated by external validation sets.

Model name	Description of external validation sets	Performance Metrics			
		Q	SE	SP	AUC
A_BBB_I	The BBB external validation set were collected from Shen's work[1], which included 246 compounds (155 BBB+ and 91 BBB- compounds).	0.882	0.981	0.714	0.978
A_HIA_I	In total, 634 oral drugs, which were not contained in the HIA training set, were collected from the DrugBank database and composed of an external validation set.[1]	0.893	0.893	---	---
CYP1A2	CYP1A2 from PubChem AID 410, CYP2C9 from PubChem AID 883, CYP2C19 from PubChem AID 899, CYP2D6 from PubChem AID 891, and CYP3A4 from PubChem AID 884 and 885. Inhibitors: PubChem Activity score equal 40 to 100; non-Inhibitors: PubChem Activity score equal 0. The detailed description about the external validation sets was given in reference [13].	0.680	89.4	0.552	0.814
CYP2C9		0.866	94.5	0.608	0.854
CYP2D6		0.803	88.4	0.583	0.841
CYP2C19		0.878	94.7	0.584	0.880
CYP3A4		0.749	83.8	0.535	0.783
AMES_Model	The external validation set contained 614 mutagens and 117 nonmutagens[28].	0.573	99.5	0.927	0.924
Biodegradation	The external validation set contained 27 novel chemicals [16].	75.0	91.3	88.9	---

Table S5. Comparison of Overall Statistics of Models in admetSAR with Previous Published Models.

Model Name	Model Description	Performance Metrics			
		Q	SE	SP	AUC
A_Caco2_I (admetSAR)	In total, 674 compounds were collected, including 303 Coca2+ and 371 Coca2- compounds. If a compound with the Caco-2 permeability value (Papp) ≥ 8×10^{-6} cm/s, it is labeled as high Caco-2 permeability, otherwise it is labeled as moderate-poor permeability.[9]	0.746	0.696	0.787	0.822
The's Caco-2 permeability classification model[9]	Constitutional descriptors	0.781	0.824	0.782	---
	Charage & molecular properties descriptors	0.810	0.851	0.808	---
	2D Autocrrelation descriptors	0.773	0.770	0.803	---
	Getaway	0.796	0.797	0.814	---
	All	0.839	0.838	0.861	---
A_PgpI_I (admetSAR)	In total, 1273 compounds were collected from Chen's work[11], including 797 Pgp inhibitors and 476 Pgp non-inhibitors.	0.786	0.872	0.641	0.853
A_PgpI_II (admetSAR)	In total, 1275 compounds were collected from Broccatelli's work[12], including 666 Pgp inhibitors and 609 Pgp non-inhibitors.	0.866	0.871	0.860	0.922
Chen's Pgp inhibitor classification model[11]	MP	---	0.771	0.696	---
	MPtECFP_4	---	0.824	0.723	---
	MPtEPFP_4	---	0.686	0.759	---
	MPtFCFP_4	---	0.835	0.732	---
	MPtFPFP_4	---	0.686	0.866	---
	MPtLCFP_4	---	0.803	0.741	---
	MPtLPFP_4	---	0.755	0.723	---
	MPtECFP_6	---	0.787	0.804	---
	MPtEPFP_6	---	0.707	0.732	---
	MPtFCFP_6	---	0.835	0.732	---
	MPtFPFP_6	---	0.782	0.804	---
	MPtLCFP_6	---	0.750	0.759	---
	MPtLPFP_6	---	0.814	0.741	---
	MPtFCFP_4	---	0.812	0.813	---
T_AMES_I (admetSAR)	In total, 8445 Compounds including 4912 AMES toxic chemicals and 3533 non AMES toxic chemicals were collected from four published papers[19-21].	0.851	0.883	0.808	0.908
Hansen's AMES model[19]	SVM	---	---	---	0.86
	GP	---	---	---	0.84
	Random	---	---	---	0.73
	*k*NN	---	---	---	0.79
M_CYP1A2I_I (admetSAR)	In total, 14903 compounds, including 7415 inhibitors and 7488 noninhibitors were collected from Cheng's work[13]. A	0.815	0.799	0.831	0.815

	compound was assigned as a CYP inhibitor if the AC_{50} (the compound concentration leads to 50% of the activity of an inhibition control) value was <10 μM, and it was considered as a non-inhibitor if AC_{50} was >57 μM. In addition, a compound was regarded as a CYP inhibitor if it has the PubChem activity score between 40 and 100, and as a noninhibitor if it has PubChem activity score equal to 0.				
Vasanthanathan' s CYP1A2 inhibitor classification model[29]	SVM^{E}	---	---	0.70	---
	RF	---	---	0.73	---
	kNN	---	---	0.68	---
	C4.5/J48	---	---	0.67	---
T_hERG_I (admetSAR)	In total, 368 molecules including 79 strong hERG inhibitors (pIC50> 6.0 mol/L)and 289 weak hERG inhibitors (pIC50≤6.0 mol/L) were collected from Marchese Robinson et al[17].	0.870	0.494	0.972	0.820
T_hERG_II (admetSAR)	In total, 806 molecules including 433 hERG inhibitors (IC50> 50 μM) and 373 hERG non-inhibitors (pIC50≤50 μM) were collected from Wang's work[18].	0.784	0.783	0.786	0.849
Su's hERG classification model[30]	raw	0.61	0.64	0.60	---
	select1289	0.68	0.53	0.70	---
	select1000	0.74	0.43	0.77	---
	Select900	0.82	0.41	0.86	---
T_TPT_I (admetSAR)	In total, 1571 compounds, including 1217 high Tetrahymena Pyriformis Toxicity (TPT) compounds and 354 low TPT compounds, were collected from Cheng's work[24]. If a compound with the $Pigc_{50}$ (the negative logarithm of 50% growth inhibitory concentration) > -0.5 was assigned as TPT, otherwise as non-TPT.	0.917	0.958	0.776	0.956
Xue's TPT classification model[31]	SVM	0.889	0.944	0.729	---
	SVM_RFE	0.904	0.935	0.820	---
R_T_TPT_I (admetSAR)	In total, 1571 compounds with $pIGC_{50}$ (ug/L) value against Tetrahymena pyriformis were collected from Cheng's work [24] for model development.	R^2= 0.761, RMSE=0.256			
Su's TPT regression model[30]	6 terms	R^2=0.695 (test set1), R^2=0.552 (test set2)			
	102 terms (abs(loadings))>0.01	R^2=0.817 (test set1), R^2=0.613 (test set2)			
	204 terms max (abs(loadings))>0.001	R^2=0.832 (test set1), R^2=0.620 (test set2),			
R_T_RAT_I (admetSAR)	In total, 10207 molecules with LD_{50} (mg/L) against rat were collected from Zhu's work [26] for rat acute toxicity regression model development.	R^2= 0.613, RMSE=0.324			
Zhu's rat acute toxicity	kNN	R^2= 0.66			

regression models[26]	RF	R^2= 0.70
	Hierarchical clustering	R^2= 0.41
	NN	R^2= 0.24
	FDA MDL QSAR	R^2= 0.29
	TOPKAT	R^2= 0.35

References

(1) Shen, J.; Cheng, F.; Xu, Y.; Li, W.; Tang, Y. Estimation of ADME properties with substructure pattern recognition. *J. Chem. Inf. Model.* **2010,** *50*, 1034-1041.

(2) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002,** *42*, 1273-1280.

(3) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011,** *3*, 33.

(4) Corinna, C.; Vladimir, V. Support-Vector Networks. *Mach. Learn.* **1995,** *20*, 273-297.

(5) V. Vapnik; S. Golowich; Smola., a. A. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, Cambridge, MA,. MIT Press. **1997,** 281–287

(6) Chang, C. C.; Lin., C.-J. LIBSVM : a library for support vector machines. <<http://www.csie.ntu.edu.tw/~cjlin/libsvm>> (Access Date: May. 18, **2011**).

(7) SMOLA A J. Regression estimation with support vector learning machines. Munchen, Master thesis, Technische University Munchen. **1996**.

(8) Ting, F. W.; Chin, J. L.; Ruby, C. W. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **2004,** *5*, 975-1005.

(9) The, H. P.; Gonzalez Alvarez, I.; Bermejo, M.; Sanjuan, V. M.; Centelles, I.; Garrogues, T. M.; Cabrera Perez, M. A. In Silico prediction of Caco-2 cell permeability by a classification QSAR approach. *Mol. Inf.* **2011,** *30*, 376-385.

(10) Wang, Z.; Chen, Y.; Liang, H.; Bender, A.; Glen, R. C.; Yan, A. P-glycoprotein substrate models using support vector machines based on a comprehensive data set. *J. Chem. Inf. Model.* **2011,** *51*, 1447-1456.

(11) Chen, L.; Li, Y.; Zhao, Q.; Peng, H.; Hou, T. ADME evaluation in drug discovery. 10. Predictions of P-glycoprotein inhibitors using recursive partitioning and naive Bayesian classification techniques. *Mol. Pharm* **2011,** *8*, 889-900.

(12) Broccatelli, F.; Carosati, E.; Neri, A.; Frosini, M.; Goracci, L.; Oprea, T. I.; Cruciani, G. A novel approach for predicting P-glycoprotein (ABCB1) inhibition using molecular interaction fields. *J. Med. Chem.* **2011,** *54*, 1740-1751.

(13) Cheng, F.; Yu, Y.; Shen, J.; Yang, L.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Classification of Cytochrome P450 Inhibitors and non-Inhibitors using Combined Classifiers. *J. Chem. Inf. Model.* **2011,** *51*, 996-1011.

(14) Cheng, F.; Yu, Y.; Zhou, Y.; Shen, Z.; Xiao, W.; Liu, G.; Li, W.; Lee, P. W.; Tang, Y. Insights into molecular basis of cytochrome p450 inhibitory promiscuity of compounds. *J. Chem. Inf. Model.* **2011,** *51*, 2482-2495.

(15) Carbon-Mangels, M.; Hutter, M. C. Selecting Relevant Descriptors for Classification by Bayesian Estimates: A Comparison with Decision Trees and Support Vector Machines Approaches for Disparate Data Sets. *Mol. Inf.* **2011,** *30*, 885 - 895.

(16) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012,** *52*, 655-669.

(17) Robinson, R. M.; Glen, R. C.; Mitchell, J. B. Development and comparison of hERG blocker classifiers: assessment on different datasets yields markedly different results. *Mol. Inf.* **2011,** *30*, 443-458.

(18) Wang, S.; Li, Y.; Wang, J.; Chen, L.; Zhang, L.; Yu, H.; Hou, T. ADMET evaluation in drug discovery. 12. Development of binary classification models for prediction of hERG potassium channel blockage. *Mol. Pharm.* **2012,** *9*, 996-1010.

(19) Hansen, K.; Mika, S.; Schroeter, T.; Sutter, A.; ter Laak, A.; Steger-Hartmann, T.; Heinrich, N.; Muller, K. R.

Benchmark data set for in silico prediction of Ames mutagenicity. *J. Chem. Inf. Model.* **2009,** *49*, 2077-2081.

(20) Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. *J. Chem. Inf. Comput. Sci.* **2004,** *44*, 1402-1411.

(21) Kazius, J.; Nijssen, S.; Kok, J.; Back, T.; Ijzerman, A. P. Substructure mining using elaborate chemical representation. *J. Chem. Inf. Model.* **2006,** *46*, 597-605.

(22) Lagunin, A.; Filimonov, D.; Zakharov, A.; Xie, W.; Huang, Y.; Zhu, F.; Shen, T.; Yao, J.; Poroikov, V. Computer-aided prediction of rodent carcinogenicity by PASS and CISOC-PSCT. *QSAR Comb. Sci.* **2009,** *28*, 806-810.

(23) Cheng., F.; Shen, J.; Li, W.; Lee, P. W.; Tang, Y. In silico prediction of terrestrial and aquatic toxicities for organic chemicals. *Chin. J. Pesti. Sci.* **2010,** *12*, 477-488.

(24) Cheng, F.; Shen, J.; Yu, Y.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In silico prediction of Tetrahymena pyriformis toxicity for diverse industrial chemicals with substructure pattern recognition and machine learning methods. *Chemosphere.* **2011,** *82*, 1636-1643.

(25) Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of reliable aqueous solubility models and their application in druglike analysis. *J. Chem. Inf. Model.* **2007,** *47*, 1395-1404.

(26) Zhu, H.; Martin, T. M.; Ye, L.; Sedykh, A.; Young, D. M.; Tropsha, A. Quantitative structure-activity relationship modeling of rat acute toxicity by oral exposure. *Chem. Res. Toxicol.* **2009,** *22*, 1913-1921.

(27) Tunkel, J.; Howard, P. H.; Boethling, R. S.; Stiteler, W.; Loonen, H. Predicting Ready Biodegradability in the Japanese Ministry of International Trade and Industry Test. *Environ. Toxicol. Chem.* **2000,** *19*, 2478-2485.

(28) Xu, C.; Cheng, F.; Chen, L.; Du, Z.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. In Silico prediction of chemical ames mutagenicity. *J. Chem. Inf. Model.* **2012**, doi: 10.1021/ci300400a.

(29) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P.; Olsen, L.; Jorgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug. Metab. Dispos.* **2009,** *37*, 658-664.

(30) Su, B. H.; Tu, Y. S.; Esposito, E. X.; Tseng, Y. J. Predictive toxicology modeling: protocols for exploring hERG classification and Tetrahymena pyriformis end point predictions. *J. Chem. Inf. Model.* **2012,** *52*, 1660-1673.

(31) Xue, Y.; Li, H.; Ung, C. Y.; Yap, C. W.; Chen, Y. Z. Classification of a diverse set of Tetrahymena pyriformis toxicity chemical compounds from molecular descriptors by statistical learning methods. *Chem. Res. Toxicol.* **2006,** *19*, 1030-1039.