

# P-glycoprotein Substrate Models Using Support Vector Machines Based on a Comprehensive Data set

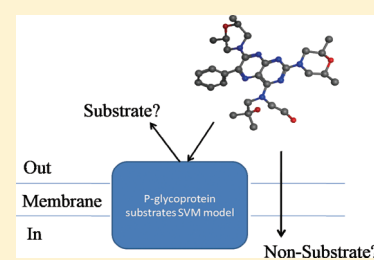
Zhi Wang,<sup>†,‡</sup> Yuanying Chen,<sup>†</sup> Hu Liang,<sup>†</sup> Andreas Bender,<sup>‡</sup> Robert C. Glen,<sup>‡</sup> and Aixia Yan<sup>\*,†</sup>

<sup>†</sup>State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, P. R. China

<sup>‡</sup>Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, United Kingdom

## Supporting Information

**ABSTRACT:** P-glycoprotein (P-gp) is one of the major ABC transporters and involved in many essential processes such as lipid and steroid transport across cell membranes but also in the uptake of drugs such as HIV protease and reverse transcriptase inhibitors. Despite its importance, reliable models predicting substrates of P-gp are scarce. In this study, we have built several computational models to predict whether or not a compound is a P-gp substrate, based on the largest data set yet published, employing 332 distinct structures. Each molecule is represented by ADRIANA.Code, MOE, and ECFP\_4 fingerprint descriptors. The models are computed using a support vector machine based on a training set which includes 131 substrates and 81 nonsubstrates that were evaluated by 5-, 10-fold, and leave-one-out (LOO) cross-validation. The best model gives a Matthews Correlation Coefficient of 0.73 and a prediction accuracy of 0.88 on the test set. Examination of the model based on ECFP\_4 fingerprints revealed several substructures which could have significance in separating substrates and nonsubstrates of P-gp, such as the nitrile and sulfoxide functional groups which have a higher frequency in nonsubstrates than in substrates. In addition structural isomerism in sugars was found to result in remarkable differences regarding the likelihood of a compound to be a substrate for P-gp.



## INTRODUCTION

The multidrug transporter P-glycoprotein (P-gp) is a member of the super family of ATP-binding cassette (ABC) transporters, which is a product of the ABCB1 (previously MDR1) gene. It consists of two homologous domains, each encompassing a hydrophobic domain and an intracellular nucleotide-binding domain (NBD); each hydrophobic domain is composed of six membrane spanning segments ( $\alpha$ -helices) separated by hydrophilic loops.<sup>1,2</sup> This protein facilitates the clearance of xenotoxins against steep concentration gradients, at the expense of ATP hydrolysis.

P-gp was first discovered in multidrug resistant cancer cells. However, it was also found in normal tissue, such as the epithelial cells of the gastrointestinal tract;<sup>3</sup> the cornea;<sup>4</sup> the biliary canaliculi front of hepatocytes; and the luminal membrane of proximal tubular epithelial cells in the kidney.<sup>5,6</sup> In the luminal membrane of the endothelial cells which line small blood capillaries and form the blood-brain barrier (BBB), blood-cerebro spinal fluid barrier (BCSFB), and blood-testis barrier, high levels of ABCB1 P-gp have been found.<sup>7–10</sup> This indicates that P-gp plays an important role in the transport of various small molecules in vital areas, such as clearing the brain of potential toxins which are P-gp substrates.<sup>11</sup>

P-Glycoprotein has attracted lots of attention in both cancer research and the pharmaceutical field because it has a significant influence on drug absorption (by transporting molecules back into the gastrointestinal (GI) lumen), distribution (by preventing

xenotoxins into important tissues like brain), metabolism (acting synergistically with cytochrome P450 3A), and excretion (by affecting both biliary and renal tubular function).<sup>12</sup> P-gp not only has an effect on absorption, distribution, metabolism, and excretion of its substrates but also induces multidrug resistance and drug–drug interactions.<sup>13</sup>

Due to the above reasons it is advisable in the drug discovery process to pay attention to the likelihood of a compound under development being transported by P-gp, since this contributes to whether a compound actually reaches its intended target (or is removed from the cell before exerting its action).

Much work investigating the structure and mechanism of P-gp has been reported.<sup>14–20</sup> It has been reported that P-gp interacts with substrates that have entered the lipid bilayer, which is different from most other transport proteins.<sup>14,15</sup> Loo and co-workers reported that the substrate binding domain appears to be located at the interface between the two transmembrane domains of P-gp and that each transmembrane domain by itself is unable to bind substrates.<sup>16</sup> Additionally it has been proven that nucleotide binding domains are not required because a truncated P-gp molecule that lacks both of these domains still retains the ability to interact with substrates.<sup>17</sup> The diversity of substrates and the complex interactions between different kinds of drugs led

Received: April 8, 2011

Published: May 23, 2011

to the hypothesis of the presence of several binding sites.<sup>18,19</sup> At least four distinct binding sites on P-gp were identified using equilibrium and kinetic radioligand binding assays. Three of them have been characterized based on the substrates (vinblastine, paclitaxel, rhodamine-123, and Hoechst33342) and modulators (XR9576, XR9051) which they are able to interact with. The fourth site is a regulatory site where elacridar and nicardipine act as modulators.<sup>20</sup> However, due to the lack of a high resolution structure of P-gp (a structure with a resolution of 3.8 Å has recently been reported<sup>19</sup>) and a lack of understanding of its dynamic behavior as a drug efflux pump at a molecular level, it is still an unsolved problem how to identify P-gp substrates reliably.

'Rules of thumb' have been introduced to characterize P-gp substrate specificity based on experimental results. Ford and Hait observed that P-gp substrates appear to be hydrophobic, with a molecular mass of 300–2000 Da.<sup>21</sup> A "rule of four" was derived by Didziapetris and co-workers which can roughly estimate whether a compound is P-gp substrate or not. This rule states that if a compound has a total of at least eight nitrogen and oxygen atoms, a molecular weight of more than 400 and an acid with  $pK_a$  of greater than four ( $(N+O) \geq 8$ ,  $MW > 400$ , and acid  $pK_a > 4$ ), it is more likely to be a P-gp substrate. In contrast, compounds with no more than four nitrogen or oxygen atoms, less than 400 molecular weight, and base  $pK_a$  of less than eight ( $(N+O) \leq 4$ ,  $MW < 400$ , and base  $pK_a < 8$ ) are likely to be nonsubstrates.<sup>22</sup> Other analyses were reported which suggest that the number and strength of hydrogen bonds are important for the interactions between P-gp and its substrates.<sup>23</sup>

Several computational methods have been developed to predict P-gp substrates, including pharmacophore methods<sup>24–26</sup> and QSAR (Quantitative Structure Activity Relationship) models. Penzotti and colleagues developed an ensemble pharmacophore model based on 195 P-gp substrates and nonsubstrates.<sup>25</sup> For the training set a prediction accuracy of 80% was achieved. However, this model did not perform well on the test set, with a prediction accuracy of 63%.<sup>25</sup> In addition, several QSAR models were reported for predicting P-gp substrates.<sup>27–34</sup> Based on the same data set including 195 P-gp substrates and nonsubstrates, De Cerqueira Lima and co-workers developed combinatorial QSAR models by the combination of several computational methods as well as several descriptor sets. The best model obtained gave a prediction accuracy of 81% on the test set.<sup>31</sup> Cabrera and co-workers reported a TOPS-MODE approach for the estimation of P-gp substrates based on 203 P-gp substrates and nonsubstrates, achieving 78% prediction accuracy on the test set.<sup>33</sup> Based on the same data set Huang and colleagues established models for prediction of P-gp substrates using a Particle Swarm (PS) algorithm and a Support Vector Machine (SVM) approach. The best model obtained in this case achieved 90% prediction accuracy on the test set.<sup>34</sup>

This study extends the above work, first by compiling a larger data set than used in previous studies<sup>25,28,29,31,33,34</sup> on the prediction of P-gp substrates. The molecules were represented initially by the descriptors calculated by ADRIANA.Code (version 2.2.2)<sup>35</sup> and MOE (2009.10)<sup>36</sup> followed by descriptor selection. In order to include interpretable structural descriptors in the modeling process extended connectivity fingerprints (ECFP) were used, calculated by Pipeline Pilot Student Edition.<sup>37</sup> Using a support vector machine (SVM), several models were generated for the identification of P-gp substrates based on different subsets of the above descriptors. For model validation 5-fold, 10-fold, and leave-one-out (LOO) cross-validation was performed on the different descriptor sets, followed

by model evaluation on an external test set. In order to gain insight into the chemical factors associated with the P-gp substrate specificity, the relative frequency of features in the substrate and nonsubstrate class was calculated and discussed in detail, based on ECFP fingerprints. Finally, correct and incorrect predictions of the model were analyzed.

## MATERIALS AND METHODS

**Data Set.** The models presented in this work were based on a data set of 332 compounds, including 206 P-gp substrates and 126 nonsubstrates (compounds are provided in SD format in the Supporting Information). All structures were originally taken from three sources: 195 compounds were collected from the work of Penzotti and colleagues<sup>25</sup> and another 91 P-gp substrates were collected from the publication of Adenot and Lahana.<sup>38</sup> In addition we obtained 257 P-gp substrates and nonsubstrates from Prof. Gerhard Ecker (Department of Medicinal Chemistry, University of Vienna).

We observed that the class labels for nine compounds were different in different data sets. (These compounds could be found in Supporting Information.) For these compounds, we use the definition that P-gp substrates are molecules which are actively transported and therefore have a higher concentration outside the cell relative to the concentration in the cytosol.<sup>20</sup> Where there was ambiguity about the molecular structure, we corrected the structures of compounds by consulting PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), Wikipedia (<http://www.wikipedia.org/>), and Chemical Information of Specialized Information Services (<http://sis.nlm.nih.gov/chemical.html>), and the structure that occurred most frequently was chosen as the 'correct' one. We employed the SDSORT program implemented in MOE<sup>36</sup> to remove duplicate compounds. The final data set of P-gp substrates and nonsubstrates contained 332 unique compounds. A Kohonen neural network<sup>39,40</sup> was used to split the data set into a training set and test set. For each neuron occupied by multiple molecules, we assigned half of the molecules into the training set and half into the test set. Those molecules occupying only one neuron were remained and projected onto a smaller sized network. This procedure was repeated until no neuron was occupied by multiple molecules. We then assigned all the unassigned molecules to the training set. This procedure may help us to obtain two subsets which can best cover the information space of the original. The data set was split into a training set (212 compounds) and a test set (120 compounds).

**Methods. Structure Representation. ADRIANA. Code Descriptors.** A total of 79 descriptors were calculated using ADRIANA.Code,<sup>35</sup> including 15 global molecular descriptors, eight size and shape descriptors, and 56 2D property autocorrelation descriptors (a full list is given in Supporting Information Table S1).

**MOE Descriptors.** The MOE<sup>36</sup> descriptors calculated comprise 185 2D descriptors, such as physicochemical properties, subdivided surface areas, atom counts and bond counts, Kier&Hall connectivity and Kappa shape indices, adjacency and distance matrix descriptors, pharmacophore feature descriptors, and partial charge descriptors. In addition, 126 3D molecular descriptors were calculated which include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors. For 3D molecular descriptors, the energy minimized conformation produced by MOE was used, and the default parameters in MOE were used to assign partial charges.

Table 1. Cross-Validation and External Validation Results for the Models Generated in This Study<sup>a</sup>

			cross-validation accuracy on training set			prediction on test set MCC/accuracy
model	number of descriptors/fingerprints		5-fold	10-fold	LOO	
Models 1 (ADRIANA. Code descriptors)	Model 1a	7	0.74	0.74	0.74	0.66/0.84
	Model 1b	6	0.74	0.74	0.75	0.61/0.82
	Model 1_rf	6	0.73	0.73	0.70	0.49/0.77
Models 2 (MOE descriptors)	Model 2a	29	0.74	0.75	0.73	0.74/0.88
	Model 2b	23	0.74	0.74	0.74	0.72/0.87
	Model 2c	19	0.73	0.74	0.74	0.62/0.83
	Model 2_rf	31	0.67	0.68	0.75	0.60/0.82
Models 3 (ADRIANA.Code and MOE descriptors)	Model 3a	32	0.74	0.74	0.75	0.68/0.85
	Model 3b	27	0.74	0.74	0.74	0.73/0.88
	Model 3c	23	0.74	0.73	0.75	0.73/0.88
	Model 3_rf	13	0.72	0.71	0.73	0.40/0.73
Model 4 (ECFP_4 fingerprints)		4583	0.70	0.70	0.69	0.60/0.82

<sup>a</sup> Models 1, 2, and 3 were generated based on physicochemical descriptors, while Model 4 employed ECFP\_4 fingerprint. "rf" denotes that descriptors used in this model were selected by the Random Forest feature selection method. It can be seen that descriptors selected by correlation analysis generally produced better models than those chosen by the Random Forest cross-validation method. Model 3c has the best prediction results on the test set with a relatively small number of descriptors; hence it was considered to be the 'best' model.

**ECFP4 Fingerprints.** Extended connectivity fingerprints (ECFP) are circular fingerprints derived using a variant of the Morgan algorithm.<sup>41</sup> The advantages of circular fingerprints are that these fingerprints can be rapidly calculated, contain a large amount of information (commonly with respect to bioactivity<sup>42,43</sup>), may represent stereochemical information, and can also be interpreted as chemical substructures. In this study, ECFP\_4 fingerprints were used.

**Descriptor Selection Methods.** Three methods were employed for descriptor selection, namely correlation analysis, Random Forest-based feature selection, and the F-score as a measure of feature relevance.

Using Pearson correlation analysis,<sup>44</sup> molecular descriptors that were not significantly correlated with activity (correlation coefficient  $r_{a0}$ ) were removed. If the pairwise correlation coefficient between any two descriptors (correlation coefficient  $r_{b0}$ ) was above a given threshold the descriptor with lower correlation with activity was removed. By adjusting different thresholds of correlation, several subsets of descriptors were selected for modeling. Two subsets of descriptors were selected by correlation analysis, namely 7 descriptors were selected when  $r_a$  and  $r_b$  were set to 0.2 and 0.9, and 6 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.8.

Random Forest cross-validation for feature selection was performed using the  $R^{45}$  implementation of Random Forests.<sup>46</sup> This function shows the cross-validated prediction performance of models with a sequentially reduced number of descriptors (ranked by descriptor importance) via a nested cross-validation procedure. In this study, 5-fold cross-validation was employed, and the number of descriptors was reduced by 10% at each step. Via this feature selection method a subset of 6 descriptors was selected, which was found to produce the minimum Random Forest cross-validation error (see Results section for details).

In addition the F-score, which measures the discrimination of two sets of real numbers, is used to calculate the importance of individual descriptors in this study. In eq 1,  $\bar{x}_i$ ,  $\bar{x}_i^{(+)}$ , and  $\bar{x}_i^{(-)}$  represent the average of the  $i^{\text{th}}$  descriptor of the whole, positive, and negative data sets;  $x_{k,i}^{(+)}$  is the  $i^{\text{th}}$  descriptor of the  $k^{\text{th}}$  positive

instance, and  $x_{k,i}^{(-)}$  is the  $i^{\text{th}}$  feature of the  $k^{\text{th}}$  negative instance; and  $n_+$  and  $n_-$  are the number of positive and negative instances, respectively. In this study, positive and negative instance refer to substrates and nonsubstrates

$$F(i) = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n_+ - 1} \sum_{k=1}^{n_+} (x_{k,i}^{(+)} - \bar{x}_i^{(+)})^2 + \frac{1}{n_- - 1} \sum_{k=1}^{n_-} (x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \quad (1)$$

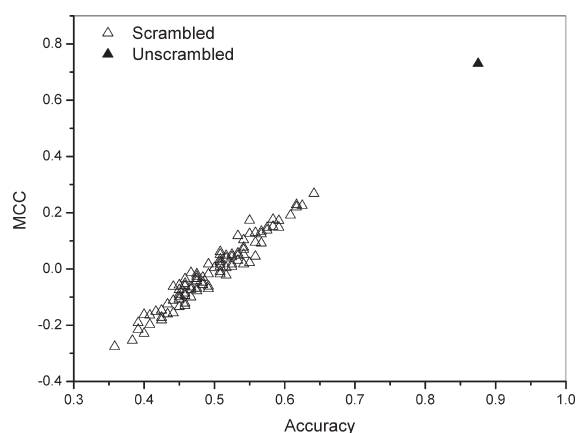
**Performance Measure of Models.** Accuracy and the Matthews Correlation Coefficient (MCC)<sup>47</sup> were employed to evaluate the predictivity of models. As can be seen in eq 2, MCC not only takes into account true positives (TP) and true negatives (TN) but also false positives (FP) and false negatives (FN). Thus, it considers different aspects of model performance, complementing the consideration of only accuracy in a classification model

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (2)$$

**Support Vector Machine.** In this work support vector machines (SVMs)<sup>48</sup> were employed to generate a classification model of P-gp substrates. SVM originated as an implementation of Vapnik's Structural Risk Minimization (SRM) principle from statistical learning theory. The special property of SVMs is that they simultaneously minimize the empirical classification error and maximize the geometric margin. Thus, SVM is also known as a maximum margin classifier.

In this study, the LIBSVM software developed by Chang and Lin was used for SVM analysis.<sup>49</sup> There are four basic kernels in LIBSVM. A radial Basis Function (RBF), which is suggested as a reasonable first choice, was used in this work.<sup>49</sup> The RBF kernel maps samples onto a higher dimensional manifold, hence it is





**Figure 1.** Matthews correlation coefficient and accuracy of Y-scrambled models and the best model (Model 3c) reported in the current work. It can be seen that performance of the model presented here is significantly superior to any of the random models, indicating that there is only a very low likelihood of chance correlations for model 3c.

able to handle nonlinear relationships between class labels and attributes.

## RESULTS AND DISCUSSION

In this work SVM models were generated initially using the full training set. Subsequently, 5-, 10-fold, and leave-one-out (LOO) cross-validation was performed. Furthermore, an external test set comprising 120 compounds was used to further evaluate the models obtained. Finally, the best model was subject to further validation by Y-scrambling in order to rule out chance correlations.

**Model 1: ADRIANA.Code Descriptors.** This part of the study was based on 79 ADRIANA.Code descriptors, where subsets of seven and six descriptors were chosen via correlation analysis and six descriptors were chosen via Random Forest which gave the lowest cross-validation error (error.cv = 0.28, see Methods section for details). Thus, three models (Model 1a, Model 1b, and Model 1\_rf) were generated using SVMs. From Table 1, one can see that Model 1a and 1b performed better on test sets than Model 1\_rf, and all three models have a cross-validation accuracy of around 0.70. Model 1a has an MCC of 0.66 and accuracy 0.84 on the test set, which is comparable in performance to Model 1b (0.60, 0.82) while outperforming Model 1\_rf (0.49, 0.77). Hence, descriptor selection based using variable correlation seems to be superior to the Random Forest based feature selection (in both cases) on the data set utilizing the descriptors employed here, independent of the precise correlation cutoff used.

**Model 2: MOE Descriptors.** Using correlation analysis, three subsets of descriptors were obtained by adjusting  $r_{a0}$  and  $r_{b0}$ . Twenty-nine descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.9; 23 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.85; and 19 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.8. Using the Random Forest cross-validation method, 31 descriptors were selected, the selection criterion being a minimum cross-validation error (here error.cv = 0.28).

Based on these four subsets of descriptors, four models (Model 2a, 2b, 2c and Model 2\_rf) were generating using SVMs. As can be seen in Table 1, the descriptors selected using correlation analysis produced better prediction models than the descriptors chosen via the Random Forest based feature selection method.

**Table 2.** Correlation Coefficient to Activity and F-Score of 36 Descriptors Used in This Study<sup>c</sup>

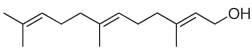
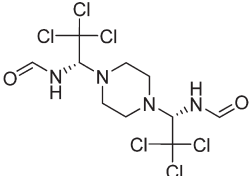
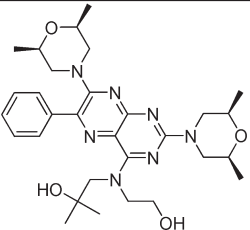
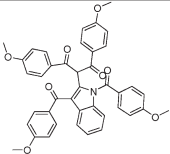
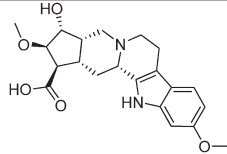
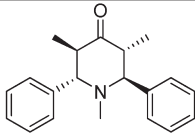
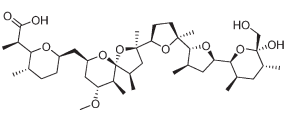
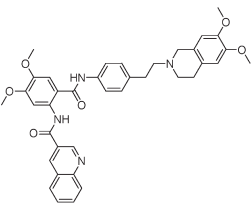
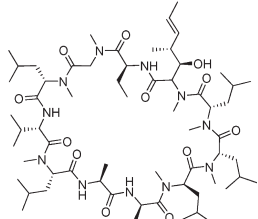
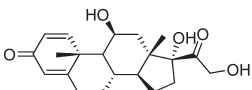
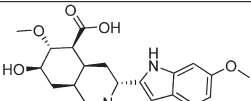
no.	descriptors	source <sup>a</sup>	R <sup>b</sup>	F-score
1	GCUT_SMR_3	M	0.32	0.11
2	BCUT_SLOGP_0	M	-0.30	0.11
3	VAdjMa	M	0.33	0.11
4	PEOE_VSA+0	M	0.32	0.11
5	std_dim2	M	0.30	0.10
6	a_hyd	M	0.34	0.10
7	opr_brigid	M	0.31	0.10
8	NAtoms	A	0.31	0.09
9	VDistEq	M	0.33	0.09
10	PEOE_RPC+	M	-0.28	0.09
11	BCUT_SMR_0	M	-0.27	0.08
12	rings	M	0.27	0.07
13	PEOE_PC-	M	-0.26	0.07
14	opr_violation	M	0.26	0.07
15	b_triple	M	-0.25	0.06
16	GCUT_PEOE_0	M	-0.25	0.06
17	SMR_VSA6	M	0.26	0.06
18	vsurf_D1	M	0.26	0.06
19	vsurf_R	M	0.24	0.06
20	2DACorr_PiChg_5	A	-0.29	0.06
21	NStereo	A	0.20	0.05
22	PEOE_RPC-	M	-0.24	0.05
23	opr_leadlike	M	-0.25	0.05
24	SlogP_VSA0	M	0.23	0.05
25	vsurf_CW1	M	-0.25	0.05
26	HDOn	A	0.21	0.04
27	balabanJ	M	-0.24	0.04
28	2DACorr_SigChg_1	A	0.21	0.04
29	PEOE_VSA+4	M	0.21	0.04
30	dens	M	-0.22	0.04
31	PEOE_VSA-0	M	0.21	0.03
32	lip_violation	M	0.20	0.03
33	2DACorr_PiChg_4	A	0.24	0.03
34	SMR_VSA5	M	0.21	0.02
35	LogS	A	-0.20	0.02
36	a_ICM	M	-0.21	0.02

<sup>a</sup> Source of descriptors: A represents descriptors calculated from ADRIANA.Code; M represents descriptors calculated from MOE. <sup>b</sup> R: correlation coefficient between descriptor and activity. <sup>c</sup> The top nine descriptors, which possess a higher F-score, all have a correlation coefficient of over 0.3 with activity. Some descriptors were found to be in agreement with known physico-chemical properties related with P-gp substrates, such as octanol/water distribution coefficient<sup>22</sup> and hydrophobicity<sup>19</sup>.

Model 2a has the best prediction performance on the test set with a MCC of 0.74 and accuracy of 0.88. As observed before, correlation-based feature selection outperforms RF based feature selection using MOE descriptors on this data set.

**Model 3: Combining ADRIANA.Code and MOE Descriptors.** In Model 3, descriptors from ADRIANA.Code and MOE were combined. By correlation analysis, three subsets of descriptors were obtained by adjusting  $r_{a0}$  and  $r_{b0}$ . 32 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.9; 27 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.85; and

Table 3. Frequently Misclassified Compounds by the Descriptor Based Models<sup>a</sup>

 farnesol (#1)	 triforine (#2)	 bibw 22 (#3)
 NSC 653278 (#4)	 reserpinic acid (#5)	 NSC 364080 (#6)
 nigericin (#7)	 XR9576 (Tariquidar) (#8)	 cyclosporin A (#9)
 prednisolone (#10)	 methyl reserpate (#11)	

<sup>a</sup> Compounds that were wrongly predicted by at least six of the descriptor based models are shown here. Bibw 22 (#3), reserpinic acid (#5), nigericin (#7), prednisolone (#10), and methyl reserpate (#11) were wrongly predicted to be substrates by all eight models.

23 descriptors were selected when  $r_{a0}$  and  $r_{b0}$  were set to 0.2 and 0.8. Using Random Forest cross-validation 13 descriptors were found to produce the minimum Random Forest cross-validation error (error.cv = 0.28).

Using each of the four descriptor subsets a separate SVM model was generated. It can be seen (Table 1) that descriptors selected by correlation analysis produced better models. Cross-validation of all models generated in section 'Model 3' showed reliable results, with the best prediction results regarding MCC and accuracy being produced by Model 3b and Model 3c, which are 0.73 and 0.88, respectively.

Of all the models obtained, Model 3c was considered to be the 'best' model given its predictive performance on the test set with a relatively small number of descriptors. To further validate Model 3c, Y-scrambling was performed.

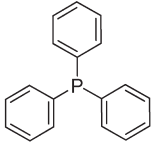
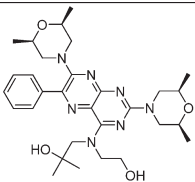
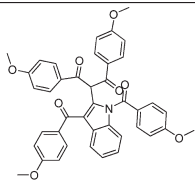
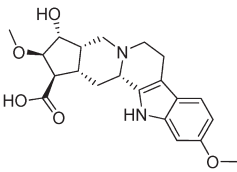
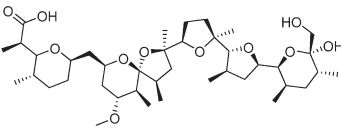
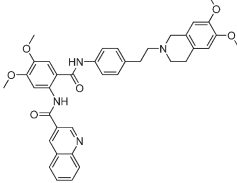
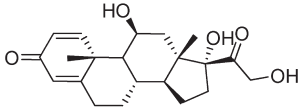
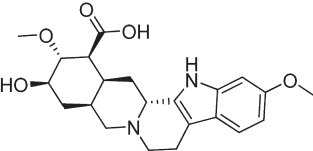
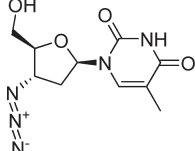
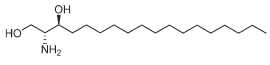
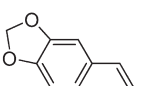
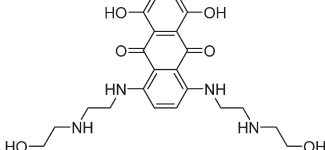
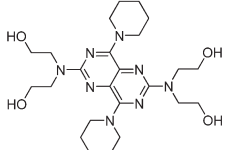
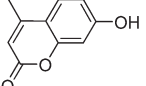
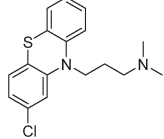
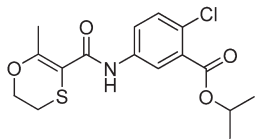
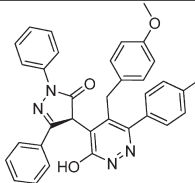
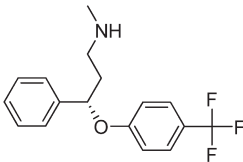
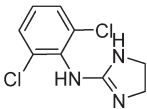
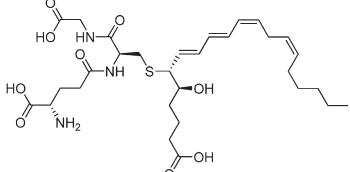
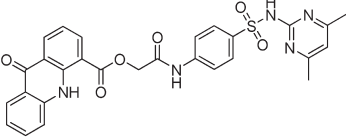
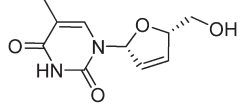
**Y-Scrambling.** Y-Scrambling<sup>50</sup> was investigated to validate that our best model (Model 3c) was not a result of chance correlation. In order to perform Y-scrambling the class label of the training set which Model 3c was based on was randomly permuted. Based on the permuted data set, a SVM model was regenerated with the same parameter settings as Model 3c, and it was used to predict the original test set. This procedure was repeated 100 times, and the resulting MCC and accuracy are visualized in Figure 1. The results from y-scrambled models are

statistically in line with what is expected on the basis of purely random prediction with the best result for the scrambling models being an MCC value of 0.27 and accuracy around 0.64. Compared to the model representing the performance of the actual best-performing model (0.73, 0.88) it can be ruled out that this model has resulted from chance correlations (at least on this data set).

**Interpretation of Feature Relevance.** As the descriptors selected by correlation analysis performed better than those from Random Forest, only the descriptors chosen by correlation analysis were investigated here. Seven descriptors from ADRI-ANA.Code and 29 from MOE were selected because of their higher correlation with activity. In order to further investigate the role of the 36 descriptors, the F-score of each descriptor was calculated which is listed in Table 2.

It can be seen (Table 2) that the top nine descriptors, which possess a higher F-score, all have a correlation coefficient greater than 0.3 with activity, indicating that these properties may contribute to resolving which molecules are P-gp substrates. It can be seen that GCUT\_SMR\_3, which is related to molar refractivity, has a higher correlation with activity and also has a higher F-score. This is also true for the descriptors BCUT\_SLOGP\_0, VAdjMa, and PEOE\_VSA+0 which represent the octanol/water distribution coefficient, molecular vertex adjacency information, and partial charge based on

Table 4. Misclassified Compounds by the ECFP\_4 Model<sup>a</sup>

<sup>a</sup> Most compounds shown in Table 3 were also wrongly predicted by ECFP\_4 model, which are Bibw 22 (#3), reserpine acid (#5), NSC 364080 (#6), nigericin (#7), R9576 (tarquidar) (#8), prednisolone (#10), and methyl reserpate (#11) in Table 3.

van der Waals surface area. This is in line with previous observations since it has been reported that the octanol/water distribution coefficient is an important factor regarding P-gp substrate-likeness.<sup>22</sup>

As can be observed from the X-ray structure of apo P-gp,<sup>19</sup> P-gp possesses large hydrophobic binding sites and binds substrates through a combination of the hydrophobic effect, aromatic interactions, and electrostatic attraction. Thus, descriptor

a\_hyd, which presents the hydrophobic pharmacophore feature of a molecule, shows a relatively high F-score, as does the descriptor rings, which counts the total number of rings and which is correlated both with hydrophobic properties as well as aromaticity of a molecule. Other descriptors of the ones selected may not be readily interpretable; however, they appear to increase the prediction power of models for P-gp substrate when being included in the model.

**Discussion of False Predictions.** Table 3 reports compounds that were wrongly predicted by at least six of our descriptor based models based on features selected by correlation analysis. Farnesol (#1), triforine (#2), and NSC 364080 (#6) were only correctly predicted by Model 1a and 1b which were based on ADRIANA.Code descriptors but always overpredicted to be substrates by the other models. NSC 653278 (#4) was also wrongly predicted to be substrates by six models and only correctly predicted by Model 2a and 2b.

XR9576 (tariquidar) (#8) was wrongly predicted to be a substrate by seven models in this study, and it is only correctly predicted by Model 1a when descriptor 2DACorr\_SigChg\_1 was used. Cyclosporin A (#9) was also wrongly predicted to be a nonsubstrate by seven models (It is probably a result of the similarity between the structure of cyclosporin A and cyclosporin D, which is in the training set, and cyclosporin D was considered as a nonsubstrate as it cannot be transported by P-gp.). However, cyclosporin A is correctly predicted by the ECFP\_4 fingerprint model. This is likely the case as the models rely on the computed property similarity of the compounds, and this alone could not discriminate cyclosporin A and D.

Bibw 22 (#3), reserpinic acid (#5), nigericin (#7), prednisolone (#10), and methyl reserpate (#11) were wrongly predicted to be substrates by all eight models. Also, it can be seen in Table 4 that these five compounds are still wrongly predicted by the ECFP\_4 model (showing the limitation of models in this study). A possible rationalization of this observation (based on molecular similarity and metabolism arguments) is that of the five compounds consistently overpredicted, reserpinic acid (#5) and methyl reserpate (#11) were reported to be substrates when they are esterified.<sup>51,52</sup> Prednisolone is likely overpredicted due to its similarity to other steroids which are indeed substrates of P-gp.

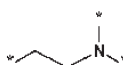
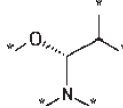
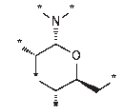
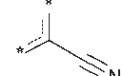
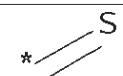
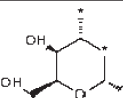
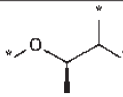
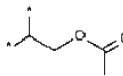
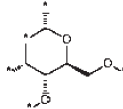
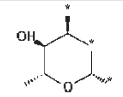
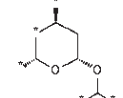
**Model 4: ECFP\_4 Fingerprints.** Model 4 was built based on 4583 ECFP\_4 fingerprints that were calculated from all 322 structures. The 5-, 10-fold, and leave-one-out (LOO) cross-validation accuracy results of Model 4 are 0.70, 0.70, and 0.69, and prediction results on test set have a MCC of 0.60 and accuracy of 0.82 (Table 1). Model 4 appears to show reliable cross-validation results and prediction on the test set.

Using structural fingerprints as an alternative method to represent molecular structure, there were some differences in which compounds were wrongly predicted by Model 4 (the ECFP\_4 model), and these are listed in Table 4. It can be seen that most compounds shown in Table 3 (false predictions in the other models) were also not correctly predicted by the ECFP\_4 model, as were NSC 364080 (#6) and XR9576 (tariquidar) (#8) in Table 3. Cyclosporin A (see #9 in Table 3) on the other hand were correctly predicted by the ECFP\_4 model.

Although the predictive power of the ECFP\_4 model is slightly weaker than descriptor models, it is possible to analyze the substructures that are selected from the model to provide a structural insight into the character of P-gp substrates.

To better understand P-gp substrates and nonsubstrates, the F-score of each ECFP\_4 substructure and the frequency of corresponding substructures among substrates and nonsubstrates were calculated. In Table 5, several representative substructures were selected (For a full list, see the sd file in Supporting Information named *ecfp4\_features\_fscore\_frequency.sd*). The first three substructures shown possess an F-score greater than 0.1, which ranks their importance in clarifying P-gp substrates from nonsubstrates. Interestingly,

**Table 5. Representative Substructures Selected from All ECFP\_4 Substructures along with F-Score and Relative Frequency among Substrates and Nonsubstrates<sup>a</sup>**

No.	Substructure	F-score	substrates_freq (%)	nonsubstrates_freq (%)
1		0.11	36.89	8.73
2		0.1	0	13.49
3		0.1	0	13.49
4		0.08	0.49	13.49
5		0.08	0.49	13.49
6		0.05	0	7.94
7		0.04	10.68	0.79
8		0.04	0	5.56
9		0.04	0	5.56
10		0.03	7.28	0
11		0.03	6.8	0

<sup>a</sup> See the main text for details.

one can see that the tertiary amine moiety (No. 1) is often contained in P-gp substrates with a frequency of 36.89% among substrates versus a frequency of 8.73% among nonsubstrates. If an alkoxy group is added next to a tertiary nitrogen (No. 2 and 3), this appears to change the substrates to nonsubstrates. It can also be seen that the nitrile group (No. 4) and thial group (No. 5) have a higher frequency in nonsubstrates than that in substrates. Based on the tetrahydro-2H-pyran group, substructures 6 and 9 were only included in nonsubstrates of our data set, while substructure 10 and 11 only appeared in substrates. It can be seen that stereoisomerism of sugar rings also leads to significant differences between substrates and nonsubstrates (substructures 7 and 8).

**Comparison to Previous Studies.** The comparison of the models obtained here with previous work<sup>25,28,29,31,33,34</sup> for



**Table 6.** Overview of the Data Sets and Model Performance Obtained in Previous Studies Analyzing P-gp Substrates<sup>a</sup>

model	data set (number of compounds)		test set		
	training set	test set	Se	Sp	A
Penzotti et al. <sup>25</sup>	144	51	0.530	0.790	0.63
Gombar et al. <sup>29</sup>	95	58	0.940	0.780	0.86
Xue et al. <sup>28</sup>	74	25	0.840	0.670	0.80
De Cerqueira Lima et al. <sup>31</sup>	144	51	0.780	0.840	0.81
Cabrera et al. <sup>33</sup>	163	40	0.820	0.720	0.78
Huang et al. <sup>34</sup>	163	40	0.910	0.890	0.90
Model 3c in this study	212	120	0.960	0.730	0.88

<sup>a</sup> Se, Sp, and A represent sensitivity, specificity, and accuracy, respectively. Se = TP/(TP+FN), Sp = TN/(TN+FP), TP (true positive), FN (false negative), TN (true negative), FP (false positive).

identifying P-gp substrates are listed in Table 4. Table 4 lists the size of each data set, the prediction accuracy for the substrates of P-gp as represented by Se (sensitivity) for the test set, and the prediction accuracy for the nonsubstrates of P-gp as represented by Sp (specificity) for the test set as well as the prediction accuracy as represented by A (accuracy) for the whole test set. The Se and Sp were calculated according to the following equations:  $Se = TP/(TP+FN)$ ,  $Sp = TN/(TN+FP)$ , where TP represents true positive, FN represents false negative, TN represents true negative, and FP represents false positive.

Due to the different sizes of the data sets (and also nonidentical structures in each set) we have avoided a direct comparison between these models. However, it can be seen that the model presented in the current work is based on a larger data set than previous models, which may extend the chemical space of prediction models, given that the predictive ability of the best models reported here are comparable to most previous models.

## CONCLUSIONS

Due to the absence of a high resolution X-ray structure of P-gp and missing knowledge on the binding mode of substrates, factors which affect P-gp specificity are still not very clear. In this case, computational methods may be useful to identify whether a compound is a P-gp substrate or not.

In this study, we compiled a larger data set than in previous studies for predicting P-gp substrates, including 332 P-gp substrates and nonsubstrates. Using physicochemical descriptors and ECFP<sub>4</sub> fingerprints, several P-gp substrate prediction models were developed. Some descriptors used in this study were found to be in agreement with known factors related to P-gp substrates, such as the octanol/water distribution coefficient<sup>22</sup> and hydrophobicity.<sup>19</sup> By investigating the F-score and frequency of ECFP<sub>4</sub> fingerprints, several substructures were found to play important roles in substrates and nonsubstrates of P-gp. For example, a tertiary amine was often contained in P-gp substrates, while a nitrile group and a sulfoxide have a higher frequency in nonsubstrates than that in substrates. Also, structural isomers of sugar rings lead to significant differences in clarifying whether a molecule is recognized as a substrate of P-gp or not.

All the models obtained were validated by cross-validation and an additional, external test set. The best model in this study has MCC of 0.73 and an accuracy of 0.88 on the test set which likely renders it of sufficient performance to be used in practical settings.

## ASSOCIATED CONTENT

**S Supporting Information.** Descriptors calculated from ADRIANA.Code were listed in Table S1. Listing of descriptors used to build models in this study that includes (a) descriptor name, (b) source of descriptors, (c) descriptions, and (d) whether these descriptors are used to build each model in this study (Table S2). In addition, a listing of the intercorrelations between the 36 descriptors and activity is provided (Table S3). All compounds used in this study can be found in sd files named `pgp_trainingset_212.sdf` and `pgp_testset_120.sdf`. Nine compounds with different definitions from different source were provided in sd file named `same_structure_different_label.sdf`. ECFP<sub>4</sub> substructures, F-score, and the frequency of corresponding substructures in substrates and non-substrates are provided in an sd file named `ecfp4_features_fscore_frequency.sdf`. Descriptors calculated for each compound are provided. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## AUTHOR INFORMATION

### Corresponding Author

\*Phone: +86-10-64421335. Fax: +86-10-64416428. E-mail: [aixia\\_yan@yahoo.com](mailto:aixia_yan@yahoo.com) or [yanax@mail.buct.edu.cn](mailto:yanax@mail.buct.edu.cn).

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (20605003 and 20975011) and the Scientific Research Foundation of Graduate School (09Li001) of Beijing University of Chemical and Technology. We thank Molecular Networks GmbH, Erlangen, Germany for making the programs ADRIANA.Code and SONNIA available for our scientific work. We thank Prof. Prof. Gerhard Ecker of The Department of Medicinal Chemistry, University of Vienna for providing their data set. Zhi Wang thanks the China Scholarship Council's for financial support for studying abroad. Robert C. Glen and Andreas Bender thank Unilever for funding.

## REFERENCES

- (1) Loo, T. W.; Clarke, D. M. Membrane Topology of A Cysteine-less Mutant of Human P-Glycoprotein. *J. Biol. Chem.* **1995**, *270*, 843–848.
- (2) Kast, C.; Canfield, V.; Levenson, R.; Gros, P. Membrane Topology of P-Glycoprotein as Determined by Epitope Insertion: Transmembrane Organization of the N-terminal Domain of MDR3. *Biochemistry* **1995**, *34*, 4402–4411.
- (3) Muller, M. B.; Keck, M. E.; Binder, E. B.; Kresse, A. E.; Hagemeyer, T. P.; Landgraf, R.; Holsboer, F.; Uhr, M. ABCB1 (MDR1)-Type P-Glycoproteins at the Blood–Brain Barrier Modulate the Activity of the Hypothalamic–Pituitary–Adrenocortical System: Implications for Affective Disorder. *Neuropsychopharmacology* **2003**, *28*, 1991–1999.
- (4) Devault, A.; Gros, P. Two Members of the Mouse MDR Gene Family Confer Multidrug Resistance with Overlapping but Distinct Drug Specificities. *Mol. Cell. Biol.* **1990**, *10*, 1652–1663.
- (5) Thiebaut, F.; Tsuruo, T.; Hamada, H.; Gottesman, M. M.; Pastan, I.; Willingham, M. C. Immunohistochemical Localization in Normal Tissues of Different Epitopes in the Multidrug Transport Protein P170: Evidence for Localization in Brain Capillaries and Cross-reactivity of One Antibody with A Muscle Protein. *J. Histochem. Cytochem.* **1989**, *37*, 159–164.
- (6) Demeule, M.; Labelle, M.; Régina, A.; Berthelet, F.; Béliveau, R. Isolation of Endothelial Cells from Brain, Lung, and Kidney: Expression



of the Multidrug Resistance P-Glycoprotein Isoforms. *Biochem. Biophys. Res. Commun.* **2001**, *281*, 827–834.

(7) Fromm, M. F. P-glycoprotein: A Defense Mechanism Limiting Oral Bioavailability and CNS Accumulation of Drugs. *Int. J. Clin. Pharmacol. Ther.* **2000**, *38*, 69–74.

(8) Cordon-Cardo, C.; O'Brien, J. P.; Casals, D.; Rittman-Grauer, L.; Biedler, J. L.; Melamed, M. R.; Bertino, J. R. Multidrug-Resistance Gene (P-Glycoprotein) is Expressed by Endothelial Cells at Blood–Brain Barrier Sites. *Proc. Natl. Acad. Sci. U.S.A.* **1989**, *86*, 695–698.

(9) Wijnholds, J.; de Lange, E. C.; Scheffer, G. L.; van den Berg, D. J.; Mol, C. A.; van der Valk, M. A.; Schinkel, A. H.; Scheper, R. J.; Breimer, D. D.; Borst, P. Multidrug Resistance Protein 1 Protects the Choroid Plexus Epithelium and Contributes to the Blood–Cerebrospinal Fluid Barrier. *J. Clin. Invest.* **2000**, *105*, 279–285.

(10) Tamai, I.; Tsuji, A. Transporter-Mediated Permeation of Drugs Across the Blood-Brain Barrier. *J. Pharm. Sci.* **2000**, *89*, 1371–1388.

(11) Lin, J. H. How Significant is the Role of P-Glycoprotein in Drug Absorption and Brain Uptake? *Drugs Today (Barc)* **2004**, *40*, 5–22.

(12) Bansal, T.; Akhtar, N.; Jaggi, M.; Khar, R. K.; Talegaonkar, S. Novel Formulation Approaches for Optimising Delivery of Anticancer Drugs Based On P-Glycoprotein Modulation. *Drug Discovery Today* **2009**, *14*, 1067–1074.

(13) Aszalos, A. Drug–Drug Interactions Affected by the Transporter Protein, P-Glycoprotein (ABCB1, MDR1) I. Preclinical Aspects. *Drug Discovery Today* **2007**, *12*, 833–837.

(14) Raviv, Y.; Pollard, H. B.; Bruggemann, E. P.; Pastan, I.; Gottesman, M. M. Photo-Sensitised Labelling of A Functional Multidrug Transporter in Living Drug-Resistant Tumour Cells. *J. Biol. Chem.* **1990**, *265*, 3975–3980.

(15) Homolya, L.; Hollo, Z.; Germann, U. A.; Pastan, I.; Gottesman, M. M.; Sarkadi, B. Fluorescent Cellular Indicators are Extruded by the Multidrug Resistance Protein. *J. Biol. Chem.* **1993**, *268*, 21493–21496.

(16) Loo, T. W.; Clarke, D. M. Superfolding of the Partially Unfolded Core-Glycosylated Intermediate of Human P-Glycoprotein into the Mature Enzyme is Promoted by Substrate-Induced Transmembrane Domain Interactions. *J. Biol. Chem.* **1998**, *273*, 14671–14674.

(17) Loo, T. W.; Clarke, D. M. The Transmembrane Domains of the Human Multidrug Resistance P-Glycoprotein are Sufficient to Mediate Drug Binding and Trafficking to the Cell Surface. *J. Biol. Chem.* **1999**, *274*, 24759–24765.

(18) Martin, C.; Berridge, G.; Mistry, P.; Higgins, C.; Charlton, P.; Callaghan, R. Drug Binding Sites On P-Glycoprotein are Altered by ATP Binding Prior to Nucleotide Hydrolysis. *Biochemistry* **2000**, *39*, 11901–11906.

(19) Aller, S. G.; Yu, J.; Ward, A.; Weng, Y.; Chittaboina, S.; Zhuo, R.; Harrell, P. M.; Trinh, Y. T.; Zhang, Q.; Urbatsch, I. L.; Chang, G. Structure of P-Glycoprotein Reveals A Molecular Basis for Poly-Specific Drug Binding. *Science* **2009**, *323*, 1718–1722.

(20) Colabufo, N. A.; Berardi, F.; Cantore, M.; Contino, M.; Inglese, C.; Niso, M.; R., P. Perspectives of P-Glycoprotein Modulating Agents in Oncology and Neurodegenerative Diseases: Pharmaceutical, Biological, and Diagnostic Potentials. *J. Med. Chem.* **2010**, *53*, 1883–1897.

(21) Ford, J. M.; Hait, W. N. Pharmacology of Drugs that Alter Multidrug Resistance in Cancer. *Pharmacol. Rev.* **1990**, *42*, 155–199.

(22) Didziapetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification Analysis of P-Glycoprotein Substrate Specificity. *J. Drug Targeting* **2003**, *11*, 391–406.

(23) Ecker, G.; Huber, M.; Schmid, D.; Chiba, P. The Importance of A Nitrogen Atom in Modulators of Multidrug Resistance. *Mol. Pharmacol.* **1999**, *56*, 791–796.

(24) Pajeva, I.; Wiese, M. Pharmacophore Model of Drugs Involved in P-Glycoprotein Multidrug Resistance: Explanation of Structural Variety (Hypothesis). *J. Med. Chem.* **2002**, *45*, 5671–5686.

(25) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuys, P. D. J. A Computational Ensemble Pharmacophore Model for Identifying Substrates of P-Glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737–1740.

(26) Stouch, T. R.; Gudmundsson, O. Progress in Understanding the Structure-Activity Relationships of P-Glycoprotein. *Adv. Drug Delivery Rev.* **2002**, *54*, 315–328.

(27) Ekins, S.; Kim, R. B.; Leake, B. F.; Dantzig, A. H.; Schuetz, E. G.; Lan, L. B.; Yasuda, K.; Shepard, R. L.; Winter, M. A.; Schuetz, J. D. Three-Dimensional Quantitative Structure-Activity Relationships of Inhibitors of P-Glycoprotein. *Mol. Pharmacol.* **2002**, *61*, 964–973.

(28) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-Glycoprotein Substrates by A Support Vector Machine Approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1497–1505.

(29) Gombar, V. K.; Polli, J. W.; Humphreys, J. E.; Wring, S. A.; Serabjit-Singh, C. S. Predicting P-Glycoprotein Substrates by A Quantitative Structure-Activity Relationship Model. *J. Pharm. Sci.* **2004**, *93*, 957–968.

(30) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach. *J. Chem. Inf. Model.* **2005**, *45*, 750–757.

(31) de Cerqueira Lima, P.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Tropsha, A. Combinatorial QSAR Modeling of P-Glycoprotein Substrates. *J. Chem. Inf. Model.* **2006**, *46*, 1245–1254.

(32) Crivori, P.; Reinach, B.; Pezzetta, D.; Poggese, I. Computational Models for Identifying Potential P-Glycoprotein Substrates and Inhibitors. *Mol. Pharmaceutics* **2006**, *3*, 33–44.

(33) Cabrera, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. A Topological Substructural Approach for the Prediction of P-Glycoprotein Substrates. *J. Pharm. Sci.* **2006**, *95*, 589–606.

(34) Huang, J.; Ma, G.; Muhammad, I.; Cheng, Y. Identifying P-Glycoprotein Substrates Using A Support Vector Machine Optimized by A Particle Swarm. *J. Chem. Inf. Model.* **2007**, *47*, 1638–1647.

(35) ADRIANA.Code; Molecular Networks GmbH: Erlangen, Germany. <http://www.molecular-networks.com> (accessed May 18, 2011).

(36) MOE (*The Molecular Operating Environment*), Version 2009.10; software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7.

(37) *Pipeline Pilot 6.1 Student ed.*, Scitegic Inc., 9665 Chesapeake Dr., Suite 9401, San Diego, CA 92123, USA.

(38) Adenot, M.; Lahana, R. J. Blood-Brain Barrier Permeation Models: Discriminating between Potential CNS and Non-CNS Drugs Including P-Glycoprotein Substrates. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 239–248.

(39) Kohonen, T. Self-Organized Formation of Topologically Correct Feature Maps. *Biol. Cybern.* **1982**, *43*, 59–69.

(40) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999; pp 81–99.

(41) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

(42) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(43) Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.

(44) Rodgers, J. L.; Nicewander, W. A. Thirteen ways to look at the correlation coefficient. *Am. Stat.* **1988**, *42*, 59–66.

(45) R Development Core Team. R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2006; ISBN 3-900051-07-0.

(46) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.

(47) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442–451.

(48) Vapnik, V.; Chapelle, O. Bounds On Error Expectation for Support Vector Machines. *Neural Comput.* **2000**, *12*, 2013–2036.

(49) Chang C. C.; Lin C. J. LIBSVM: a library for support vector machine. 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm> (accessed May 18, 2011).

- (50) Rücker, C.; Rücker, G.; Meringer, M.  $\gamma$ -Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.
- (51) Yasuda, K.; Lan, L. B.; Sanglard, D.; Furuya, K.; Schuetz, J. D.; Schuetz, E. G. Interaction of Cytochrome P450 3A Inhibitors with P-Glycoprotein. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 323–332.
- (52) Polli, J. W.; Wring, S. A.; Humphreys, J. E.; Huang, L.; Morgan, J. B.; Webster, L. O.; Serabjit-Singh, C. S. Rational Use of In Vitro P-Glycoprotein Assays in Drug Discovery. *J. Pharmacol. Exp. Ther.* **2001**, *299*, 620–628.