Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents

Y. Xue,[†,‡,§] Z. R. Li,[§] C. W. Yap,[†] L. Z. Sun,[†] X. Chen,[†] and Y. Z. Chen*[,†]

Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore-MIT Alliance, E-04-10, 4 Engineering Drive 3, Singapore, 117576, and Department of Chemistry, Sichuan University, Chengdu, 610064, P. R. China

Received April 19, 2004

Statistical-learning methods have been developed for facilitating the prediction of pharmacokinetic and toxicological properties of chemical agents. These methods employ a variety of molecular descriptors to characterize structural and physicochemical properties of molecules. Some of these descriptors are specifically designed for the study of a particular type of properties or agents, and their use for other properties or agents might generate noise and affect the prediction accuracy of a statistical learning system. This work examines to what extent the reduction of this noise can improve the prediction accuracy of a statistical learning system. A feature selection method, recursive feature elimination (RFE), is used to automatically select molecular descriptors for support vector machines (SVM) prediction of P-glycoprotein substrates (P-gp), human intestinal absorption of molecules (HIA), and agents that cause torsades de pointes (TdP), a rare but serious side effect. RFE significantly reduces the number of descriptors for each of these properties thereby increasing the computational speed for their classification. The SVM prediction accuracies of P-gp and HIA are substantially increased and that of TdP remains unchanged by RFE. These prediction accuracies are comparable to those of earlier studies derived from a selective set of descriptors. Our study suggests that molecular feature selection is useful for improving the speed and, in some cases, the accuracy of statistical learning methods for the prediction of pharmacokinetic and toxicological properties of chemical agents.

INTRODUCTION

In the study of pharmacodynamic, pharmacokinetic, and toxicological properties of drugs and other chemical agents, a variety of molecular descriptors has been developed and routinely used for describing physicochemical and structural properties of chemical agents.[1−7] These descriptors were initially developed for the construction of quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSPR) of structurally related compounds.[8] They have been extensively used for the statistical-learning-based prediction of pharmacodynamic, pharmacokinetic, and toxicological properties of chemical agents including drug-likeness,[9−11] blood-brain barrier penetration,[12,13] human intestinal absorption,[4] drug-receptor binding,[14−16] drug metabolism,[17] cellular membrane partitioning,[18] chemical space navigation,[19] and antibacterial activity.[20,21]

Some of these molecular descriptors are developed for the study of a particular type of properties of a group of structurally related chemical agents. Thus these descriptors may not be universally applicable for other agents or for the prediction of other properties. For instance, descriptors for the QSAR of relatively small sets of related agents are not applicable for the analysis of chemical diversity.[22] The use of descriptors unrelated to a particular type of properties or biological activity likely generates noise in a statistical learning system, which may affect the prediction accuracy of that system.[22] In some cases, it is difficult to manually select descriptors useful for a particular property. Thus methods capable of automatic selection of molecular descriptors are desirable. The redundancy in molecular descriptors can be partially reduced by means of feature selection methods.[23−27] Feature selection methods have been found to increase the prediction accuracy of statistical learning classification of some systems. Examples include the prediction of drug activities,[23] cancer tissue sample classification using microarray data,[24] gene selection for cancer classification,[25] and splice site prediction.[26,27] It is thus of interest to examine whether feature selection methods can be explored for automatic selection of molecular descriptors and for improvement of the prediction accuracy of pharmacodynamic, pharmacokinetic, and toxicological properties of chemical agents by statistical learning methods.

In this work, a widely used feature selection method is used to automatically select the molecular descriptors for the prediction of three different pharmacokinetic and toxicological properties of chemical agents. One is the prediction of P-glycoprotein (P-gp) substrates, which facilitates early identification and elimination of drug candidates of low efficacy or high potential of multidrug resistance.[28−31] This is a process that only involves active transport via binding to P-gp. The second is the prediction of human intestinal absorption (HIA) of chemical agents, an important indicator for drug absorption.[32−35] HIA primarily involves passive transport with a small portion of compounds being absorbed

* Corresponding author phone: 65-6874-6877; fax: 65-6774-6756; e-mail: yzchen@cz3.nus.edu.sg.
† National University of Singapore.
‡ Singapore-MIT Alliance.
§ Sichuan University.

10.1021/ci049869h CCC: $27.50
Published on Web 07/10/2004

Table 1. Molecular Descriptors and Their Classes Used in This Work[a]

descriptor class	no. of descriptors in class	descriptors
simple molecular properties	18	molecular weight, number of ring structures, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, element counts
molecular connectivity and shape	28	molecular connectivity indices, valence molecular connectivity indices, molecular shape kappa indices, kappa alpha indices, flexibility index
electrotopological state	84	electrotopological state indices and atom type electrotopological state indices
quantum chemical properties	13	atomic charge on the most positively charged H atom, largest negative charge on an non-H atom, polarizability index, hydrogen bond acceptor basicity (covalent HBAB), hydrogen bond donor acidity (covalent HBDA), molecular dipole moment, absolute hardness, softness, ionization potential, electron affinity, chemical potential, electronegativity index, electrophilicity index
geometrical properties	16	molecular size vectors (distance of the longest separated atom pairs, combined distance of the longest separated three atoms, combined distance of the longest separated four atoms), molecular van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, polar molecular surface area, sum of solvent accessible surface areas of positively charged atoms, sum of solvent accessible surface areas of negatively charged atoms, sum of charge weighted solvent accessible surface areas of positively charged atoms, sum of charge weighted solvent accessible surface areas of negatively charged atoms, sum of van der Waals surface areas of positively charged atoms, sum of van der Waals surface areas of negatively charged atoms, sum of charge weighted van der Waals surface areas of positively charged atoms, sum of charge weighted van der Waals surface areas of negatively charged atoms

[a] The total number of descriptors is 159.

by active transport through various transporters. The third is the prediction of compounds that induce torsades de pointes (TdP), an uncommon adverse drug reaction responsible for the withdrawal of some marketed drugs.[36−38] A substantial portion of TdP is due to channel blocking, but other unknown mechanisms are also involved. The different mechanisms of these three problems make them useful for testing feature selection methods. The computed results are further compared to those of earlier studies to examine whether our selected descriptors are capable of giving similar or better classification performance with respect to those derived from a preselected set of descriptors.

The feature selection method used in this work is the recursive feature elimination (RFE) method, which has recently gained popularity due to its effectiveness for discovering informative features or attributes in drug activity analysis and cancer tissue sample classification.[23,25] Support vector machine (SVM)[39,40] is used in this work as the statistical learning method for the prediction of the three pharmacokinetic and toxicological properties of chemical agents. SVM has been applied to a wide range of pharmacological and biomedical problems including drug-likeness,[9−11] drug blood-brain barrier penetration prediction,[41] drug-receptor binding,[14] and drug metabolism.[17] In many cases SVM has been found to be consistently superior to other supervised learning methods[10,12,42−44] and less sensitive to overfitting.[25] Thus SVM is an appropriate platform to evaluate the effectiveness of feature selection methods in improving the accuracy of statistical learning methods for the prediction of pharmacodynamic, pharmacokinetic, and toxicological properties of chemical agents.

METHODS

Selection of Data Sets. P-gp substrates are collected from the literature[28−33] that are either described as being transported by P-gp or reported to induce overexpression of P-gp thereby directly contributing to MDR. Nonsubstrates of P-gp are those specifically described as not transportable by P-gp. A total of 116 substrates and 85 nonsubstrates of P-gp are collected.

Chemical agents absorbable (HIA+) or nonabsorbable (HIA−) by human intestine are from those described in the literature, in which the "measured absorption rate" of 70% is used as the criterion for dividing chemical agents into HIA+ and HIA− classes.[33,34] A total of 131 HIA+ and 65 HIA− agents are collected. In general, a relatively smaller number of agents with low intestinal absorption is specifically reported in the literature.[35] Thus, the number of known HIA+ agents are expected to be significantly larger than those of HIA− agents.

Eighty-five TdP inducing (TdP+) agents are collected from ArizonaCERT,[45] Micromedex,[46] Drug Information Handbook,[47] and Meyler's side effects of drugs.[48] Those involved in QT prolongation without information about their effect on TdP are not included. Two hundred seventy-six non-TdP causing (TDP−) agents are obtained from the search of Micromedex, Drug Information Handbook, and American Hospital Formulary Service (AHFS)[49] for agents with no reported case of TdP.

Molecular Descriptors. The molecular descriptors used in this work are selected from those commonly used in the literature.[7] These descriptors are first screened manually to remove those that are apparently redundant or irrelevant to the pharmacokinetic and toxicological properties. A total of 159 descriptors are selected, as given in Table 1, which can be divided into five classes based on their properties. There are 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 84 descriptors in the class of electrotopological state, 13 descriptors in the class of quantum chemical properties, and 16 descriptors in the class of geometrical properties. These descriptors are computed from the 2D and 3D structure of each agent using our own designed molecular descriptor computing program.[50] The 3D structure of each agent is

generated from its 2D structure by using Concord v4.02 software.

Simple descriptors are counts of special atoms and chemical bonds in the molecules. Examples of these descriptors include the number of ring structures, number of rotatable bonds, number of hydrogen bond donors and acceptors, molecular weight, and element counts. Molecular connectivity chi indices and shape Kappa indices encode information about molecular size, shape, branching, unsaturation, heteroatom content, and cyclicity.[51,52] The electrotopological state indices are numerical values computed for each atom in a molecule, which encode information about both the topological environment of that atom and the electronic interactions due to all other atoms in the molecule.[53,54] Quantum chemical descriptors are used to describe electrostatic and electronic properties of a molecule. These descriptors are calculated using molecular orbital energies and wave functions of electronic motion in a molecule, which can be obtained by solving the Schrödinger equation of electronic motion.[55] The computed quantum chemical descriptors include partial atomic charges, the highest occupied and lowest unoccupied molecular orbital energies, dipole moment, polarizability, and other descriptors derived from them.[6,56] Geometric descriptors encode the 3D-structural features of a molecule. These include van der Waals volume, solvent accessible surface area, molecular surface area, van der Waals surface area, and the corresponding quantities associated with partial charges and polarity etc.[57,58]

All the P-gp substrates and nonsubstrates, HIA+ and HIA− agents, and TdP+ and TdP− agents used in this study are available as Supporting Information. The 159 descriptors for each compound are also provided as Supporting Information.

SVM Algorithm. The theory of SVM has been extensively described in the literature.[39,40] Thus only a brief description is given here. SVM is based on the structural risk minimization (SRM) principle from statistical learning theory.[39] In linearly separable cases, SVM constructs a hyperplane which separates two different classes of vectors with a maximum margin. In this case, a vector corresponds to a chemical agent, and this vector is represented by $\mathbf{x}_i$, with structural and physicochemical descriptors of the chemical agent as its components. This is done by finding another vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ and satisfies the following conditions

$$\mathbf{w}\cdot\mathbf{x}_i + b \geq +1, \text{ for } y_i = +1 \text{ class 1 (positive samples)} \quad (1)$$

$$\mathbf{w}\cdot\mathbf{x}_i + b \leq -1, \text{ for } y_i = -1 \text{ class 2 (negative samples)} \quad (2)$$

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin, and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. After the determination of $\mathbf{w}$ and b, a given vector $\mathbf{x}_i$ can be classified by

$$sign[(\mathbf{w}\cdot\mathbf{x}) + b] \quad (3)$$

In nonlinearly separable cases, SVM maps the input variable into a high-dimensional feature space using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. An example of a kernel function is the Gaussian kernel which has been extensively used in different studies with good results.[41,43,59]

$$K(\mathbf{x}_i,\mathbf{x}_j) = e^{-||\mathbf{x}_j-\mathbf{x}_i||^2/2\sigma^2} \quad (4)$$

Linear support vector machine is applied to this feature space and then the decision function is given by

$$f(\mathbf{x}) = sign(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x},\mathbf{x}_i) + b) \quad (5)$$

where the coefficients α_i^0 and b are determined by maximizing the following Lagrangian expression

$$\sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i,\mathbf{x}_j) \quad (6)$$

under the following conditions:

$$a_i \geq 0 \text{ and } \sum_{i=1}^{l}\alpha_i y_i = 0 \quad (7)$$

A positive or negative value from eq 3 or eq 5 indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

As in the case of all discriminative methods,[60,61] the performance of SVM classification can be measured by the quantity of true positives *TP*, true negatives *TN*, false positives *FP*, false negatives *FN*, sensitivity $SE = TP/(TP+FN)$ which is the prediction accuracy for positive examples in this work, and specificity $SP = TN/(TN+FP)$ which is the prediction accuracy for negative examples in this work. The overall prediction accuracy (Q) and Matthews correlation coefficient (C)[62] are also used to measure the prediction accuracies and can be given by

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$C = \frac{TP*TN - FN*FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \quad (9)$$

Feature Selection Method. Features refer to descriptors used by statistical learning methods for classification of specific problems. Feature selection methods have been introduced for the improvement of classification performance of statistical learning methods and for the selection of features meaningful in discriminating two data sets.[22−27] One approach, the recursive feature elimination (RFE) method, has gained popularity due to its effectiveness for discovering informative features or attributes in cancer classification and drug activity analysis.[23,25] Thus in this work, the RFE method is used.

It has been suggested that the ranking criterion for feature selection can be based on the change in the objective function upon removing each feature.[63] To improve the efficiency of training, this objective function is represented by a cost function J for the ith feature computed by using training set only. When a given feature is removed or its weight w_i is reduced to zero, the change in the cost function $DJ(i)$ is given by

$$DJ(i) = \frac{1}{2}\frac{\partial^2 J}{\partial w_i^2}(Dw_i)^2 \quad (10)$$

The case of $Dw_i = w_i - 0$ corresponds to the removal of feature *i*.

Guyon et al. have used RFE to reduce the number of descriptors of a linear SVM classification system for cancer detection from gene selection data.[25] In the corresponding linear SVM classifier, the cost function is $J = (1/2)||w||^2 - \alpha^T \mathbf{l}$, where **1** is an *m* dimensional identity vector (*m* is the number of compounds in the training set). Therefore $DJ(i) = (1/2)\, w_i^2$ and w_i^2 can be used as a feature ranking criterion. Yu et al. have used RFE to reduce the descriptors of a nonlinear SVM classification system of polynomial kernels for prediction of drug activity.[23] However, because of the diversity and complexity of chemical agents, the use of linear and polynomial kernels may not always be sufficient for accurate prediction of various pharmaceutical and biological properties. Thus, in this work, SVM classification systems of Gaussian kernels are used. In this case, the cost function to be minimized (under the constraints $0 \le \alpha\, k \le C$ and $\sum_k \alpha_k y_k = 0$) is

$$J = (1/2)\alpha^T \mathbf{H}\alpha - \alpha^T \mathbf{1} \quad (11)$$

where **H** is the matrix with elements $y_i\, y_j \exp(-||x_i - x_j||^2/(2\sigma^2))$, and **1** is an *m* dimensional identity vector (*m* is the number of compounds in training set).

To compute the change in cost function caused by removing input component *i*, the parameters α's are kept unchanged and the matrix **H** is recomputed. The resulting ranking coefficient is

$$DJ(i) = (1/2)\alpha^T \mathbf{H}\alpha - (1/2)\alpha^T \mathbf{H}(-i)\alpha \quad (12)$$

where $\boldsymbol{H}(-i)$ is the matrix computed by using the same method as that of matrix $\boldsymbol{H}$ but with its *i*th component removed. One or more of features with the smallest $DJ(i)$ can thus be eliminated.

Computation Procedure. The computation procedure in this work is outlined as the following: The SVM classification system for this study was trained by using a Gaussian kernel function. The training was conducted by sequential variation of the parameter σ in the special region against the whole training data set. The prediction accuracy of this SVM system during the training process was evaluated by means of 5-fold cross-validation. In the first step, for a fixed σ, the SVM classifier is trained by using the complete set of features (molecular descriptors) described in the previous section. The second step is to compute the ranking criterion score $DJ(i)$ for each feature in the current set by using eq 12. All of the computed $DJ(i)$ is subsequently ranked in descending order. The third step is to remove the *m* features with smallest criterion scores. In this work, *m* was chosen to be 5 as that used in earlier studies.[27] In the fourth step, the SVM classification system is retrained by using the remaining set of features, and the corresponding prediction accuracy is computed by means of 5-fold cross-validation. The first to fourth steps are then repeated for other values of σ. After the completion of these procedures, the set of features and parameter σ that give the best prediction accuracy are selected.

Table 2. SVM and SVM+RFE Prediction Accuracy of the Substrates and Nonsubstrates of P-Glycoprotein by Using 5-Fold Cross-Validation

		substrates			nonsubstrates				
method	cross-validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)	C
SVM	1	14	10	58.3	9	7	56.3	57.5	0.14
	2	15	2	88.2	11	5	68.8	78.8	0.58
	3	24	14	63.2	10	4	71.4	65.4	0.31
	4	14	5	73.7	14	4	77.8	75.7	0.51
	5	11	7	61.1	14	7	66.7	64.1	0.28
	average			**68.9**			**68.2**	**68.3**	**0.37**
SVM+RFE	1	17	7	70.8	12	4	75.0	72.5	0.45
	2	15	2	88.2	11	5	68.8	78.8	0.58
	3	30	8	78.9	13	1	92.9	82.7	0.65
	4	15	4	78.9	15	3	83.3	81.1	0.62
	5	16	2	88.9	16	5	76.2	82.1	0.65
	average			**81.2**			**79.2**	**79.4**	**0.59**

Table 3. SVM and SVM+RFE Prediction Accuracy of the Human Intestinal Absorption (HIA+) and Nonabsorption (HIA−) of Chemical Agents by Using 5-Fold Cross-Validation

		HIA+			HIA−				
method	cross-validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)	C
SVM	1	22	5	81.5	7	5	58.3	74.4	0.40
	2	18	3	85.7	8	3	72.7	81.3	0.58
	3	37	3	92.5	7	5	58.3	84.6	0.54
	4	16	4	80.0	7	8	46.7	65.7	0.28
	5	18	5	78.3	12	3	80.0	79.0	0.57
	average			**83.4**			**63.2**	**77.0**	**0.48**
SVM+RFE	1	22	5	81.5	10	2	83.3	82.1	0.61
	2	20	1	95.2	11	0	100.0	96.9	0.93
	3	35	5	87.5	8	4	66.7	82.7	0.53
	4	18	2	90.0	10	5	66.7	80.0	0.59
	5	22	1	95.7	13	2	86.7	92.1	0.83
	average			**90.0**			**80.7**	**86.7**	**0.70**

Table 4. SVM and SVM+RFE Prediction Accuracy of TdP Inducing (TdP+) Agents and Non-TdP Causing (TdP−) Agents Using 5-Fold Cross-Validation

		TdP+				TdP−			
method	cross-validation	TP	FN	SE (%)	TN	FP	SP (%)	Q (%)	C
SVM	1	6	11	35.3	45	4	91.8	77.3	0.33
	2	10	6	62.5	42	6	87.5	81.3	0.50
	3	12	7	63.2	63	4	94.0	87.2	0.61
	4	8	5	61.5	46	3	93.9	87.1	0.59
	5	10	10	50.0	54	9	85.7	77.1	0.36
	average			**54.5**			**90.6**	**82.0**	**0.48**
SVM+RFE	1	8	9	47.1	45	4	91.8	80.3	0.44
	2	13	3	81.3	41	7	85.4	84.4	0.62
	3	15	4	78.9	59	8	88.1	86.1	0.63
	4	10	3	76.9	46	3	93.9	90.3	0.71
	5	10	10	50.0	55	8	87.3	78.3	0.39
	average			**66.8**			**89.3**	**83.9**	**0.56**

RESULTS AND DISCUSSION

Effect of Feature Selection on Classification Accuracy. The prediction accuracy of SVM classification systems using the RFE method (termed as SVM+RFE) and those without using RFE (termed as SVM) is evaluated by means of 5-fold cross-validation method. The computed accuracies for each fold and the average accuracies of P-gp substrates and nonsubstrates are given in Table 2, those of HIA+ and HIA− agents are given in Table 3, and those of TdP+ and TdP− agents are given in Table 4, respectively. The corresponding overall prediction accuracy (*Q*) and Matthews correlation coefficient (*C*) are also given in Tables 2−4.

The average accuracy for the SVM prediction of P-gp substrate and P-gp nonsubstrates is 68.9% and 68.2%, that

Table 5. Distribution of the Molecular Descriptors in the Reduced Set Selected by the RFE Method[a]

system	total number of descriptors in the reduced set	number of descriptors in descriptor class	percentage in each class (%)	descriptor class
P-gp		11	50.0	electrotopological state
		4	18.2	quantum chemical
	22	3	13.6	connectivity and shape
		2	9.1	geometric
		2	9.1	simple molecular properties
HIA		13	48.1	electrotopological state
		7	25.9	connectivity and shape
	27	3	11.1	quantum chemical
		3	11.1	geometric
		1	3.7	simple molecular properties
TdP		17	54.8	electrotopological state
		6	19.4	quantum chemical
	31	5	16.1	connectivity and shape
		3	9.7	geometric
		0	0.0	simple molecular properties

[a] The total number of descriptors in the original data set is 159.

for the prediction of HIA+ and HIA− agents is 83.4% and 63.2%, and that for the prediction of TdP+ and TdP− agents is 54.5% and 90.6%, respectively. By using RFE, the total number of descriptors is significantly reduced from 159 to 22 for P-gp, to 27 for HIA, and to 31 for TdP. The average accuracies for the prediction of P-gp and HIA are substantially improved by using each of these reduced set of descriptors, respectively. These are 81.2% and 79.2% for P-gp substrate and P-gp nonsubstrates and 90.0% and 80.7% for HIA+ and HIA− agents, respectively. On the other hand, the average accuracy for the prediction of TdP remains at the same level as that without using RFE, which is 66.8% for and 89.3% for the prediction of TdP+ and TdP− agents, respectively. One possible reason for the insensitivity of the prediction accuracy with respect to feature selection is that TdP involves multiple mechanisms,[37,38] which is likely a more dominant factor for affecting prediction accuracy than descriptor redundancy. Our study seems to suggest that RFE is useful for removing redundant descriptors, which helps to increase the computational efficiency of statistical learning system. In some cases, the feature selection method RFE is capable of improving the accuracy of SVM classification of pharmacokinetic behavior of chemical agents.

Comparison with Other Classification Studies. The effect of feature selection on classification performance can be further evaluated by comparison with other classification studies of the same systems that use preselected descriptors. Direct comparison between our results and those from other studies may not be appropriate because of differences in the use of data set, descriptors, evaluation, and classification methods. For instance, our study of SVM classification of P-gp substrates shows that evaluation based on 5-fold cross-validation can be different from that based on the use of a more evenly represented training set and an independent evaluation set. Nonetheless, a tentative comparison may provide some crude estimate regarding the approximate level of accuracy of our method with respect to those obtained by other studies that used more selective descriptors.

The P-gp substrate prediction accuracy of 81.2% by using SVM+RFE is substantially improved with respect to the value of 63% derived from the ensemble pharmacophore model that uses a selective set of hydrophobe and hydrogen bond descriptors.[31] The reported accuracies of HIA+ predictions are 77%−87% by using partitioned total surface models,[64] 80% by using neural network method together with 2D topological descriptors,[65] and 97% by using SAR models together with physicochemical and structural descriptors.[66] The reported accuracy for HIA− prediction is 85% by using SAR models.[66] Our prediction accuracy of 90.0% for HIA+ and 80.7% for HIA− by using SVM+RFE is thus comparable to the results from these methods that use selective sets of descriptors.

There has been no other reported study of direct computational prediction of TdP-causing risk. Thus our results are tentatively compared to those of the prediction of QT prolongation, which frequently but not necessarily lead to TdP.[67] Agents that induce QT prolongation usually cause disruption of the outward potassium currents by blocking potassium ion channels, particularly the HERG K+ channel, which might then induce TdP.[68] There is however no definitive correlation between QT prolongation and TdP.[67,69] For instance, verapamil causes QT prolongation but does not induce TdP, whereas procainamide and disopyramide cause TdP but are not potent inhibitors of the HERG K+ channel.[69] Our prediction accuracies of 66.8% for TdP+ and 89.3% for TdP− are comparable to the values of 71% for QT prolongation and 93% for non-QT prolongation derived by the use of Ghose and Crippen descriptors.[70]

RFE Selected Molecular Descriptors. Table 5 gives the distribution of the RFE-method-selected descriptors for each of the three classification problems along with their descriptor classes. These descriptors are listed in Table 6. Descriptors from all of the classes are selected by the RFE method. Those from the class of electrotopological state constitute the largest percentage of the descriptors selected, which is consistent with a linear discriminant analysis of structure-based descriptors for multidrug resistant (MDR) agents that showed that 60% of the molecular descriptors important for MDR are topological in nature.[71] A large variety of descriptors in this class, such as those of different functional groups and hydrophobic properties, are important for characterization of pharmacodynamic, pharmacokinetic, and toxicological properties.[71,72] There are also a substantial number of descriptors from the quantum chemical, connectivity, and

Table 6. Molecular Descriptors Selected by the RFE Method for the Classification of Three Pharmacokinetic and Toxicological Properties: P-Glycoprotein Substrates (P-gp), Human Intestine Absorption (HIA), and a Rare Side-Effect Torsades de Pointes (TdP)[a]

system (primary mechanism)					
P-gp (AT)	HIA (PT)	TdP (CB)	descriptors selected	description	class
√	√	√	$^{5}\chi_{CH}$	simple molecular connectivity chi indices for cycle of 5 atoms	connectivity
√	√	√	$^{5}\chi^{v}_{CH}$	valence molecular connectivity chi indices for cycle of 5 atoms	connectivity
√	√	√	S(13)	atom-type H estate sum for CH_n (unsaturated)	electrotopological state
√	√	√	S(16)	atom-type estate sum for $-CH_3$	electrotopological state
√	√	√	S(25)	atom-type estate sum for =C<	electrotopological state
√	√	√	π_i	polarizability index	quantum chemical properties
√	√		Ndonr	number of H-bond donors	simple molecular properties
√	√		S(1)	atom-type H Estate sum for −OH	electrotopological state
√	√		S(20)	atom-type Estate sum for =CH-	electrotopological state
√		√	S(18)	atom-type Estate sum for $>CH_2$	electrotopological state
√		√	S(21)	atom-type Estate sum for: CH: (aromatic)	electrotopological state
√		√	S(36)	atom-type Estate sum for >N-	electrotopological state
√		√	q^+	atomic charge on the most positively charged H atom	quantum chemical properties
√		√	m	molecular dipole moment	quantum chemical properties
√		√	w	electrophilicity index	quantum chemical properties
√		√	dis2	length vector (longest third atom)	geometrical properties
	√	√	$^{3}\chi^{v}_{C}$	valence molecular connectivity chi indices for cluster	connectivity
	√	√	$^{6}\chi_{CH}$	simple molecular connectivity chi indices for cycle of 6 atoms	connectivity
	√	√	S(5)	atom-type H Estate sum for > NH	electrotopological state
	√	√	S(10)	atom-type H Estate sum for :CH: (sp^2, aromatic)	electrotopological state
	√	√	S(26)	atom-type Estate sum for: C:-	electrotopological state
	√	√	S(31)	atom-type Estate sum for >NH	electrotopological state
	√	√	S(35)	atom-type Estate sum for :N:	electrotopological state
	√	√	Sanc	sum of solvent accessible surface areas of negatively charged atoms	geometrical properties
	√	√	Sancw	sum of charge weighted solvent accessible surface areas of negatively charged atoms	geometrical properties
√			$^{3}\chi^{v}_{P}$	valence molecular connectivity chi indices for path order 3	connectivity
√			ncocl	count of Cl atoms	simple molecular properties
√			S_{car}	sum of Estate indices of carbon atoms	electrotopological state
√			S(9)	atom-type H Estate sum for =CH− (sp^2)	electrotopological state
√			S(12)	atom-type H Estate sum for CH_n (saturated)	electrotopological state
√			Sapcw	sum of charge weighted solvent accessible surface areas of positively charged atoms	geometrical properties
	√		dis3	length vectors (longest distance of fourth atom)	geometrical properties
	√		$^{2}\chi$	simple molecular connectivity chi index for path order 2	connectivity
	√		$^{3}\chi_{C}$	simple molecular connectivity chi indices for cluster	connectivity
	√		$^{6}\chi^{v}_{CH}$	valence molecular connectivity chi indices for cycle of 6 atoms	connectivity
	√		S(34)	atom-type Estate sum for =N-	electrotopological state
	√		S(39)	atom-type Estate sum for −OH	electrotopological state
	√		S(40)	atom-type Estate sum for =O	electrotopological state
	√		ϵ_a	hydrogen bond donor acidity (covalent HBDA)	quantum chemical properties
	√		A	electron affinity	quantum chemical properties
		√	$^{4}\chi^{v}_{PC}$	valence molecular connectivity chi indices for path/cluster	connectivity
		√	S_{hal}	sum of Estate indices of halogen atoms	electrotopological state
		√	S(2)	atom-type H Estate sum for =NH	electrotopological state
		√	S(4)	atom-type H Estate sum for $-NH_2$	electrotopological state
		√	S(29)	atom-type Estate sum for $-NH_2$	electrotopological state
		√	S(37)	atom-type Estate sum for $-N\ll(NO_2)$	electrotopological state
		√	S(41)	atom-type Estate sum for −O-	electrotopological state
		√	q^-	largest negative charge on an non-H atom	quantum chemical properties
		√	η	absolute hardness	quantum chemical properties

[a] The primary mechanism for each of these properties is given in terms of AT (active transport), PT (passive transport), and CB (channel blocking).

geometric classes. These descriptors are important for describing electrostatic, structural-framework, and geometric properties of chemical compounds.[72−74]

A number of descriptors selected by the RFE method are for more than one of the classification problems. Six of the descriptors are selected in all of the three classification systems. These describe molecular connectivity of ring structures, topological property of hydrophobic groups, and polarizability index. Thus these quantities appear to be important for describing the pharmacokinetic and toxicological properties of chemical agents studied in this work. Such a conclusion is consistent with descriptors used in the earlier studies of P-gp[71] and HIA.[72]

There are 7 additional electrotopological and quantum chemical descriptors jointly selected by the RFE method for the P-gp and TdP systems. A substantial portion of the TdP agents are channel blockers.[68] Thus, the agents for both systems are binders of membrane-bound transporter or channel, and it is not surprising that they share several additional descriptors known to be important for protein binding. Only 3 additional descriptors are jointly selected by the RFE method for the P-gp and HIA systems. These

describe the broad features of hydrogen bond, −OH and CH− groups. Unlike P-gp, which solely involves active transport, only a very small portion of HIA agents are actively transported. Thus the number of shared descriptors is expected to be less than that of the P-gp and TdP systems because of the limited diversity of actively transported HIA agents.

Many of the HIA agents are passively transported,[72] and some of the TdP agents are not channel blockers.[67,69] Thus these systems are expected to be descried by descriptors not selected for P-gp. One finds that there are 8 descriptors jointly selected by the RFE method for the HIA and TdP systems. In addition to connectivity properties of clusters, the majority of these descriptors measure polar properties. This suggests that certain electrotopological and polar features are shared in the description of passive transport and the unknown mechanisms of TdP.

There are also a number of descriptors exclusively selected for each of the problems. For instance, 6 descriptors are selected for P-gp, which describe carbon-based electrotopological properties, solvent accessible surface area for positively charged atoms, and the number of Cl atoms. These descriptors are likely selected for describing certain special P-gp substrates. There are 9 descriptors exclusively selected for HIA. These describe polar properties, molecular size, cluster connectivity, and various +N-, −OH, and =O electrotopological properties, which are likely important for describing passive transport across membranes. There are 9 descriptors exclusively selected for TdP, which describe charge property, valence connectivity, and various O, N, NH, NH2 electrotopological properties. These descriptors possibly describe binding to certain types of proteins.

From Table 6, one finds that the descriptors selected by the RFE method are primarily uncorrelated to each other. The majority of the descriptors removed by the RFE method, particularly those of electrotopological state, geometrical, and quantum chemical properties, are found to have a certain level of correlation to some of the descriptors selected. The rest of the RFE removed descriptors are mostly simple molecular properties (such as molecular weight, the number of specific types of atoms, and the number of rings), geometrical properties (such as molecular volume and surface areas), and connectivity properties (such as index for clusters and paths). These descriptors are not selected because they are not useful for distinguishing molecules in the specific data sets for the particular pharmacokinetic and toxicological property studied in this work. For instance, an examination of the molecules in all of the three pairs of data sets shows that molecules in a positive data set and those in the corresponding negative data sets are in the same range of molecular weight, volume, and surface areas.

CONCLUSION

Feature selection methods are capable of automatic selection of molecular descriptors and reduction of the noise generated by the use of overlapping and redundant molecular descriptors. This reduction appears to be helpful in enhancement of the performance of statistical learning method for the prediction of pharmacokinetic and toxicological properties of chemical agents. Recent efforts are directed at the improvement of the efficiency and speed of feature selection methods,[27] which can further help to optimally select molecular descriptors and enable the development of more accurate and efficient computational tools for the prediction of pharmacodynamic, pharmacokinetic, and toxicological properties of chemical agents.

In this work, a feature selection method is incorporated into SVM classification systems for dividing molecules into two classes according to specific pharmacokinetic or toxicological property. This method can also be applied to the prediction of pharmacokinetic and toxicological properties in a continuous fashion, i.e., the prediction of structure−property relationship. For instance, feature selection method can be combined with regression SVM[75] and regression neural network methods[65,76−78] for providing nonlinear QSPR of specific pharmacokinetic or toxicological property.

Supporting Information Available: P-gp substrates and nonsubstrates, HIA+ and HIA− agents, and TdP+ and TdP− agents and the 159 descriptors for each compound. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Karelson, M. Introduction. In *Molecular descriptors in QSAR/QSPR*; Karelson, M., Ed.; Wiley-Interscience: New York; 2000; pp 1−11.
(2) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(3) Katritzky, A. R.; Gordeeva, E. V. Traditional topological indices vs electronic, geometrical and combined molecular descriptors in QSAR/QSPR research. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 835−857.
(4) Cruciani, G.; Pastor, M.; Guba, W. Volsurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.
(5) Kier, L. B.; Hall, L. H. *Molecular structure description: The electrotopological state*; Academic Press: San Diego, 1999.
(6) Karelson, M.; et al. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027−1043.
(7) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, 2000.
(8) Katritzky, A. R.; Tatham, D. B.; Maran, U. Theoretical descriptors for the correlation of aquatic toxicity of environmental pollutants by quantitative structure-toxicity relationships. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1162−1176.
(9) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048−2056.
(10) Byvatov, E.; Fechner, U.; Sadowski, J. S. G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(11) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.
(12) Doniger, S.; Hofmann, T.; Yeh, J. Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms. *J. Comput. Biol.* **2002**, *9*, 849−864.
(13) Crivori, P.; Cruciani, G.; Carrupt, P. A.; Testa, B. Predicting blood-brain barrier permeation from three-dimensional molecular structure. *J. Med. Chem.* **2000**, *43*, 2204−2216.
(14) Bock, J. R.; Gough, D. A. A new method to estimate ligand−receptor energetics. *Mol. Cell Proteomics* **2002**, *1*, 904−910.
(15) Zamora, I.; Oprea, T.; Cruciani, G.; Pastor, M.; Ungell, A. L. Surface descriptors for protein−ligand affinity prediction. *J. Med. Chem.* **2003**, *46*, 25−33.
(16) Filipponi, E.; Cruciani, G.; Tabarrini, O.; Cecchetti, V.; Fravolini, A. QSAR study and VolSurf characterization of anti-HIV quinolone library. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 203−217.
(17) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by

human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019−2024.
(18) Alifrangis, L. H.; Christensen, I. T.; Berglund, A.; Sandberg, M.; Hovgaard, L.; Frokjaer, S. Structure−property model for membrane partitioning of oligopeptides. *J. Med. Chem.* **2000**, *43*, 103−113.
(19) Oprea, T. I.; Zamora, I.; Ungell, A. L. Pharmacokinetically based mapping device for chemical space navigation. *J. Comb. Chem.* **2002**, *4*, 258−266.
(20) Cronin, M. T. D.; Aptula, A. O.; Dearden, J. C.; Duffy, J. C.; Netzeva, T. I.; Patel, H.; Rowe, P. H.; Schultz, T. W.; Worth, A. P.; Voutzoulidis, K.; Schüürmann, G. Structure-based classification of antibacterial activity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 869−878.
(21) Aptula, A. O.; Kühne, R.; Ebert, R. U.; Cronin, M. T. D.; Netzeva, T. I.; Schüürmann, G. Modeling discrimination between antibacterial and nonantibacterial activity based on 3D molecular descriptors. *QSAR Comb. Sci.* **2003**, *22*, 113−128.
(22) Bayada, D. M.; Hamersma, H.; van Geerestein, V. J. Molecular diversity and representativity in chemical databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1−10.
(23) Yu, H.; Yang, J.; Wang, W.; Han, J. Discovering compact and highly discriminative features or feature combinations of drug activities using support vector machines. *Proceeding of the IEEE computer society bioinformatics conference (CSB)* **2003**, 220−228.
(24) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**, *16*, 906−914.
(25) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389−422.
(26) Degroeve, S.; De Baets, B.; Van de Peer, Y.; Rouzé, P. Feature subset selection for splice site prediction. *Bioinformatics* **2002**, *18*, S75−S83.
(27) Furlanello, C.; Serafini, M.; Merler, S.; Jurman, G. An accelerated procedure for recursive feature ranking on microarray data. *Neural Networks* **2003**, *16*, 641−648.
(28) Bain, L. J.; McLachlan, J. B.; LeBlanc, G. A. Structure−activity relationships for xenobiotic transport substrates and inhibitory ligands of P-glycoprotein. *Environ. Health Perspect.* **1997**, *105*, 812−818.
(29) Litman, T.; Zeuthen, T.; Skovsgaard, T.; Stein, W. D. Structure−activity relationships of P-glycoprotein interacting drugs: kinetic characterization of their effects on ATPase activity. *Biochim. Biophys. Acta* **1997**, *1361*, 159−168.
(30) Seelig, A. A general pattern for substrate recognition by P-glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.
(31) Penzotti, J. E.; Lamb, M. L.; Evensen, E.; Grootenhuis, P. D. J. A computational ensemble pharmacophore model for identifying substrates of P-glycoprotein. *J. Med. Chem.* **2002**, *45*, 1737−1740.
(32) Wessel, M. D.; Jurs, P. C.; Tolan, J. W.; Muskal, S. M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 726−735.
(33) Zhao, Y. H.; Le, J.; Abraham, M. H.; Hersey, A.; Eddershaw, P. J.; Luscombe, C. N.; Boutina, D.; Beck, G.; Sherborne, B.; Cooper, I.; Platts, J. A. Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure−activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* **2001**, *90*, 749−784.
(34) Abraham, M. H.; Zhao, Y. H.; Le, J.; Hersey, A.; Luscombe, C. N.; Reynolds, D. P.; Beck, G.; Sherborne, B.; Cooper, I. On the mechanism of human intestinal absorption. *Eur. J. Med. Chem.* **2002**, *37*, 595−605.
(35) Klopman, G.; Stefan, L. R.; Saiakhov, R. D. ADME evaluation. 2. A computer model for the prediction of intestinal absorption in humans. *Eur. J. Pharm. Sci.* **2002**, *17*, 253−263.
(36) Saunders, W. B. *Dorland's illustrated medical dictionary*; London, 2000.
(37) Layton, D.; Key, C.; Shakir, S. A. Prolongation of the QT interval and cardiac arrhythmias associated with cisapride: Limitations of the pharmacoepidemiological studies conducted and proposals for the future. *Pharmacoepidemiol. Drug Saf.* **2003**, *12*, 31−40.
(38) De Ponti, F.; Poluzzi, E.; Cavalli, A.; Recanatini, M.; Montanaro, N. Safety of nonantiarrhythmic drugs that prolong the QT interval or induce torsade de pointes: An overview. *Drug Saf.* **2002**, *25*, 263−286.
(39) Vapnik, V. N. *The nature of statistical learning theory*; Springer: New York, 1995.
(40) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.* **1998**, *2*, 127−167.
(41) Trotter, M. W. B.; Buxton, B. F.; Holden, S. B. Support vector machines in combinatorial chemistry. *Measurement Control* **2001**, *34*, 235−239.
(42) Brown, M. P. S.; Grundy, W. N.; Lin, D.; Cristianini, N.; Sugnet, C.; Ares, J. M.; Haussler, D. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 262−267.
(43) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(44) Cai, Y. D.; Liu, X. J.; Xu, X. B.; Chou, K. C. Support vector machines for predicting HIV protease cleavage sites in protein. *J. Comput. Chem.* **2002**, *23*, 267−274.
(45) ArizonaCERT Drugs that prolong the QT interval and/or induce torsades de pointes ventricular arrhythmia. http://www.arizonacert.org/medical-pros/drug-lists/drug-lists.htm (November 18, 2003).
(46) MICROMEDEX. *MICROMEDEX*; MICROMEDEX: Greenwood Village, CO, edition expires 12/2003.
(47) Lacy, C. F.; et al. *Drug information handbook*; Lexi-Comp, Inc.: Hudson, OH, 2002.
(48) Dukes, M. N. G. *Meyler's side effects of drugs*; Excerpta Medica: Amsterdam, 1996.
(49) Bethesda. *AHFS drug information*; American Society of Health-System Pharmacists, Inc.: 2001.
(50) Xue, Y.; Yap, C. W.; Li, Z. R.; Chen, Y. Z. Evaluation of a method for improving the computation speed of molecular descriptors for drug property analysis. *Acta Pharmacol. Sin.* 2004, Submitted for publication.
(51) Kier, L. B.; Hall, L. H. *Molecular connectivity in structure−activity analysis*; Research Studies Press: Wiley: Letchworth, Hertfordshire, England; New York, 1986.
(52) Hall, L. H.; Kier, L. B. The molecular connectivity chi indices and kappa shape indices in structure−property modeling. In *Reviews of Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol. 2, pp 367−412.
(53) Hall, L. H.; Mohney, B. K.; Kier, L. B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76−82.
(54) Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039−1045.
(55) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models, 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(56) Thanikaivelan, P.; Subramanian, V.; Raghava, J.; Rao, J. R.; Nair, B. U. Application of quantum chemical descriptors in quantitative structure activity and structure property relationship. *Chem. Phys. Lett.* **2000**, *323*, 59−70.
(57) Hopfinger, A. J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.
(58) Tsodikov, O. V.; Record, M. T. J.; Sergeev, Y. V. Novel computer program for fast exact calculation of accessible and molecular surface areas and average surface curvature. *J. Comput. Chem.* **2002**, *23*, 600−609.
(59) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct.-Act. Relat.* **2001**, *20*, 227−240.
(60) Roulston, J. E. Screening with tumor markers. *Mol. Pharmacol.* **2002**, *20*, 153−162.
(61) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412−424.
(62) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405*, 442−451.
(63) Kohavi, R.; John, G. H. Wrappers for feature subset selection. *Artificial Intelligence* **1997**, *97*, 273−324.
(64) Bergstrom, C. A.; Strafford, M.; Lazorova, L.; Avdeef, A.; Luthman, K.; Artursson, P. Absorption classification of oral drugs based on molecular surface properties. *J. Med. Chem.* **2003**, *46*, 558−570.
(65) Niwa, T. Using general regression and probabilistic neural networks to predict human intestinal absorption with topological descriptors derived from two-dimensional chemical structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 113−119.
(66) Zmuidinavicius, D.; DidziaPetris, R.; Japertas, P.; Avdeef, A.; Petrauskas, A. Classification structure−activity relations (C−SAR) in prediction of human intestinal absorption. *J. Pharm. Sci.* **2003**, *92*, 621−633.
(67) Malik, M.; Camm, A. J. Evaluation of drug-induced QT interval prolongation: implications for drug approval and labeling. *Drug Saf.* **2001**, *24*, 323−351.

(68) Vandenberg, J. I.; Walker, B. D.; Campbell, T. J. HERG K+ channels: friend and foe. *Trends Pharmacol. Sci.* **2001**, *22*, 240−246.

(69) Muzikant, A. L.; Penland, R. C. Models for profiling the potential QT prolongation risk of drugs. *Curr. Opin. Drug Discov. Devel.* **2002**, *5*, 127−135.

(70) Roche, O.; Trube, G.; Zuegge, J.; Pflimlin, P.; Alanine, A.; Schneider, G. A virtual screening method for prediction of the hERG potassium channel liability of compound libraries. *Chembiochem* **2002**, *3*, 455−459.

(71) Bakken, G. A.; Jurs, P. C. Classification of multidrug-resistance reversal agents using structure-based descriptors and linear discriminant analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(72) Egan, W. J.; Merz, K. M. J.; Baldwin, J. J. Prediction of drug absorption using multivariate statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.

(73) Klopman, G.; Shi, L. M.; Ramu, A. Quantitative structure−activity relationship of multidrug resistance reversal agents. *Mol. Pharmacol.* **1997**, *52*, 323−334.

(74) Abraham, M. H. Scales of solute hydrogen-bonding: Their construction and appplication to physicochemical and biochemical processes. *Chem. Soc. Rev.* **1993**, *22*, 73−83.

(75) Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. An accurate QSPR study of O−H bond dissociation energy in substituted phenols based on support vector machines. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 669−677.

(76) Mosier, P. D.; Jurs, P. C. QSAR/QSPR studies using probabilistic neural networks and generalized regression neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*(*6*), 1460−1470.

(77) Douali, L.; Villemin, D.; Cherqaoui, D. Neural networks: Accurate nonlinear QSAR model for HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2003**, *43*(*4*), 1200−1207.

(78) Korolev, D.; Balakin, K. V.; Nikolsky, Y.; Kirillov, E.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Nikolskaya, T. Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.* **2003**, *46*(*17*), 3631−43.

CI049869H