Combinatorial QSAR Modeling of P-Glycoprotein Substrates

Patricia de Cerqueira Lima,[†] Alexander Golbraikh,[†] Scott Oloff,[†] Yunde Xiao,[‡] and Alexander Tropsha*,[†]

Division of Natural and Medicinal Chemistry, The Laboratory for Molecular Modeling, School of Pharmacy, CB# 7360, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7360, and Targacept Inc., 200 East First Street, Suite 300, Winston−Salem, North Carolina 27101-4165

Received September 29, 2005

Quantitative structure−activity (property) relationship (QSAR/QSPR) models are typically generated with a single modeling technique using one type of molecular descriptors. Recently, we have begun to explore a combinatorial QSAR approach which employs various combinations of optimization methods and descriptor types and includes rigorous and consistent model validation (Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−95). Herein, we have applied this approach to a data set of 195 diverse substrates and nonsubstrates of P-glycoprotein (P-gp) that plays a crucial role in drug resistance. Modeling methods included *k*-nearest neighbors classification, decision tree, binary QSAR, and support vector machines (SVM). Descriptor sets included molecular connectivity indices, atom pair (AP) descriptors, VolSurf descriptors, and molecular operation environment descriptors. Each descriptor type was used with every QSAR modeling technique; so, in total, 16 combinations of techniques and descriptor types have been considered. Although all combinations resulted in models with a high correct classification rate for the training set (CCR_{train}), not all of them had high classification accuracy for the test set (CCR_{test}). Thus, predictive models have been generated only for some combinations of the methods and descriptor types, and the best models were obtained using SVM classification with either AP or VolSurf descriptors; they were characterized by $CCR_{train} = 0.94$ and 0.88 and $CCR_{test} = 0.81$ and 0.81, respectively. The combinatorial QSAR approach identified models with higher predictive accuracy than those reported previously for the same data set. We suggest that, in the absence of any universally applicable "one-for-all" QSAR methodology, the combinatorial QSAR approach should become the standard practice in QSPR/QSAR modeling.

INTRODUCTION

P-glycoprotein (P-gp) or MDR1 is a member of the superfamily of ATP-binding cassette transporter proteins. It is composed of two homologous and symmetrical cassettes, each containing six transmembrane domains that are separated by an intracellular flexible linker polypeptide loop which binds and hydrolyzes ATP.[1] P-gp has been found on the epithelial cells of the gastrointestinal tract;[2] the biliary canalicular front of hepatocytes;[2] on the apical surfaces of epithelial cells in small biliary ductules in the liver,[2] of small ductules of pancreatic ducts,[2] of the proximal tubules in the kidney,[2] and of superficial columnar cells in the colon and jejunum;[2] and on the surfaces of cells in both the cortex and medulla of the adrenal gland, in the luminal membrane of brain capillary endothelial cells,[3] and so forth. High expression of P-gp has also been observed in multidrug-resistant cancer cells.[4,5] Multidrug resistance caused by P-gp is a significant obstacle to the successful treatment of cancer patients.[4] Human P-gp plays a critical role in the success of drug candidates since it acts as a drug efflux transporter by extruding a large number of structurally diverse and mechanistically unrelated compounds out of cells.[6]

Although P-gp's physiological function and the mechanism of molecular recognition and cellular transport by the enzyme are not clearly understood, some believe that the major physiological role of the multidrug transporters is the protection of our cells and tissues against xenobiotics.[7] From a pharmacological point of view, P-gp can cause poor absorption, distribution, metabolism, and excretion of its substrates.[1,8] Evidence in the variation of interindividual and intraindividual intestinal P-gp expression may contribute to the variability in the absorption of oral drugs.[9]

Pharmacophore models for identifying P-gp substrates[10] have been described in the literature.[8,11,12] Penzotti and collaborators studied a database of 195 P-gp substrates and nonsubstrates.[11] Using the pharmacophore-based three-dimensional whole molecule descriptors,[13−16] and the CONAN software,[17] they have developed an ensemble pharmacophore model capable of filtering out substrates from nonsubstrates. The model was able to correctly classify 80% of the training set, but it showed a poor performance for the test set (prediction classification accuracy was 63%).[11]

In a different published study, Stouch and Gudmundsson developed a predictive QSAR model for a data set of 75 substrates and 23 nonsubstrates of P-gp.[12] The model had an overall prediction success rate of 75%; misclassified compounds were identified as large and conformationally flexible. The only validation method used for the model was

* Corresponding author phone: +1 (919) 966-2955; fax: +1 (919) 966-0204; e-mail: alex_tropsha@unc.edu.
[†] University of North Carolina at Chapel Hill.
[‡] Targacept Inc.

10.1021/ci0504317 CCC: $33.50
Published on Web 03/02/2006

permutation of the biological activities followed by the generation of new models; there was no division of the data set into training and test sets. Consequently, the model's external predictive power was not evaluated. In light of recent studies demonstrating the utmost importance of external validation,[18] this omission compromises the applicability of Stouch's model to the virtual screening of new P-gp compounds. Furthermore, it was not clear how the overall classification rate was calculated. Usually, it is estimated by the ratio N_{corr}/N_{tot}, where N_{tot} and N_{corr} are the total number of compounds and the number of compounds classified correctly, respectively. This ratio is a poor estimator of classification accuracy, if the sizes of different classes are significantly different, as was the case in the Stouch and Gudmundsson study.[12] For example, if one of the classes includes 80% of the compounds, assigning all of the compounds to this class will give a prediction accuracy as high as 80%. A brief review of previous published work on P-gp modeling and SAR can be found in ref 12, and none of the reported models appeared to have a high *external* prediction accuracy. Thus, the development of a validated predictive QSAR model for the classification of drug candidates as P-gp substrates or nonsubstrates remains a significant challenge.

Despite many years of research and a large variety of approaches, there exists no "gold standard" QSAR approach that guarantees the best model for every data set. Recently, we began to advance the combinatorial QSAR approach that explores various combinations of optimization methods and descriptor types and includes rigorous and consistent validation.[19] Herein, we have applied this approach to a data set of 195 diverse substrates and nonsubstrates of P-gp, studied by Penzotti et al.[11] Optimization methods included *k*-nearest neighbors (*k*NN) classification,[19] decision tree (DT),[20] binary QSAR (BQ),[20,21] and support vector machines (SVM).[22] Descriptor sets included molecular connectivity indices (MolconnZ),[23] atom pair (AP)[24,25] descriptors, VolSurf descriptors,[26] and molecular operation environment (MOE) descriptors.[20] QSAR studies were carried out separately for each method and each descriptor type. In total, 16 combinations of methods and descriptor types were used to develop QSAR models. The model classification accuracy was estimated as the average of accuracies for each class as discussed in detail below. The best models were obtained using SVM classification with AP descriptors or with VolSurf descriptors; these models yielded higher predictive accuracies than QSAR models reported previously for the same data set. Our studies emphasize that the exploratory nature of the combinatorial QSAR approach helps in identifying highly predictive models for a particular data set, whereas a conventional approach to QSAR studies using only one method and one type of descriptors has a higher chance to fail.

METHODS

Data Set. A total of 195 P-gp substrates and nonsubstrates were taken from ref 11. To compare our approach with the 3D QSAR methodology used by Penzotti and co-workers,[11] we have used the same training and test sets to build and validate models. The training set contained 144 compounds (76 substrates and 68 nonsubstrates) and the test set contained 51 compounds (32 substrates and 19 nonsubstrates). The Tanimoto index[27] of 0.18 calculated using Daylight fingerprints[28] shows that this data set is quite diverse.[11]

All chemical structures were represented in SMILES (simplified molecular input line entry system) notation.[29,30] The SMILES strings were converted to representative three-dimensional (3D) conformations using the *dbtranslate* function of the Sybyl 6.9 UNITY module.[31] These structures were used to calculate all types of 2D and 3D descriptors.

Descriptors. Descriptors are molecular properties calculated from the chemical structure (e.g., molecular weight, the number hydrogen-bond donors and acceptors, surface area, different invariants of a molecular graph, etc.). The descriptor types used in the combinatorial QSAR studies of 195 P-gp substrates and nonsubstrates are described in the following sections. Each descriptor type was used separately with each QSAR method.

MolconnZ 4.05 Descriptors. MolconnZ descriptors included valence, path, cluster, path/cluster, and chain molecular connectivity indices,[32−34] κ molecular shape indices,[35,36] topological[37] and electrotopological[38−41] state indices, differential connectivity indices,[33,42] the graph's radius and diameter,[43] Wiener[44] and Platt[45] indices, Shannon[46] and Bonchev-Trinajstić[47] information indices, counts of different vertices,[23] and counts of paths and edges between different types of vertices.[23] Descriptors were normalized by range scaling, so that they had values within the interval [0, 1]. The total number of descriptors was 381. After elimination of low variance descriptors (with variance lower than 0.05), this number was reduced to 141.

Atom Pair Descriptors. AP descriptors were calculated using the GenAP[25] program developed in our laboratory. GenAP implements an approach proposed by Carhart et al.[24] AP descriptors are defined by the types of atoms (or centers of double or triple chemical bonds) and the shortest topological distances between them. The topological distance is the number of atoms along the shortest path connecting two atoms in a molecular graph. The general form of an atom pair descriptor is as follows:

Atom type *i* ... (distance)... Atom type *j*

In this study, the following 15 SYBYL[31] atom types were used: (1) negative charge center, NCC; (2) positive charge center, PCC; (3) hydrogen-bond acceptor, HA; (4) hydrogen-bond donor, HD; (5) aromatic ring center, ARC; (6) nitrogen atoms, N; (7) oxygen atoms, O; (8) sulfur atoms, S; (9) phosphorus atoms, P; (10) fluorine atoms, FL; (11) chlorine, bromine, and iodine atoms, HAL; (12) carbon atoms, C; (13) all other elements, OE; (14) triple-bond center, TBC; and (15) double-bond center, DBC. The total number of possible pairs is 120. A total of 15 distance bins were defined in the interval from graph distance zero (i.e., zero atoms separating an atom pair) to 14 and greater. Thus, the total number of descriptors was $120 \times 15 = 1800$. Many of the AP descriptors frequently have a value of zero for all molecules in a data set (when certain atom types or atom pairs are absent in all molecular structures). For our data set of 195 compounds, the total number of AP descriptors was 1379, but after the elimination of low variance descriptors, just 173 descriptors were used in the QSAR studies.

VolSurf Descriptors. VolSurf descriptors are obtained from 3D interaction energy grid maps.[26] The calculation of

VolSurf descriptors includes the following steps. (i) A grid is built around a molecule. (ii) An interaction field (with water, dry, amide, and carbonyl probes representing solvent, hydrophobic, and hydrogen-bond acceptor and donor effects) is calculated at each grid point. (iii) Eight or more energy values are assigned, and for each energy value, the number of grid points inside the surface corresponding to this energy (*Vol*ume descriptors) or belonging to this surface (*Surf*ace descriptors) is calculated.[26]

The main advantage of VolSurf descriptors is that they are alignment-free; that is, alignment of the molecules is not necessary prior to the descriptor calculations. VolSurf descriptors include size and shape descriptors, hydrophilic and hydrophobic region descriptors, interaction energy moments, and other descriptors.[26]

The grid step was equal to 0.5 Å. A total of 72 descriptors were calculated per structure. All of them were used in the QSAR studies.

MOE Descriptors. MOE descriptors[20] include both 2D and 3D molecular descriptors. 2D descriptors include physical properties, subdivided surface areas, atom and bond counts, Kier and Hall connectivity[32−34] and κ shape indices,[35,36] adjacency and distance matrix descriptors,[43,44,48,49] pharmacophore feature descriptors, and partial charge descriptors.[20,50] 3D molecular descriptors include potential energy descriptors, surface area, volume and shape descriptors, and conformation-dependent charge descriptors.[20] MOE descriptors were range-scaled. The total number of descriptors was 189. All of them were used in the QSAR studies.

QSAR METHODS

Correct Classification Rate (CCR). Typically, CCR is defined as the ratio of compounds classified correctly to the total number of compounds. This definition of CCR has a major drawback when the number of compounds belonging to different classes are significantly different. Suppose there are two classes, and as in ref 12, 75 compounds belong to class 1 and 23 compounds belong to class 0. Assume that some hypothetical "model" will assign all of the compounds to class 1. Then CCR = 0.76, since 75/(75 + 23) = 0.76; that is, we would believe that our "model" is very good contrary to common sense.

To avoid artificial over-rating of the classification model accuracy, in this study, CCR was defined as follows. Let N be the total number of compounds in a data set and N_1 and N_0 be the number of P-gp substrates and nonsubstrates, respectively. Of course, $N_0 + N_1 = N$. Let TP and TN be the number of substrates predicted as substrates (true positives) and the number of nonsubstrates predicted as nonsubstrates (true negatives), respectively. Then

$$\mathrm{CCR} = 0.5(\mathrm{TP}/N_1 + \mathrm{TN}/N_0) \quad (1)$$

In this case, for the hypothetical example described above, we obtain CCR = 0.5, and our "model" assigning all compounds to class 1 does not seem to be more accurate than the random assignment of each molecule with a probability of 0.5 to the class of substrates or nonsubstrates. In addition to CCR, we have also assessed the classification accuracy of our models by sensitivity Se = TP/N_1 and specificity Sp = TN/N_0. Finally, we have also used the normalized enrichment values for each class as characteristics of classification accuracy:[51]

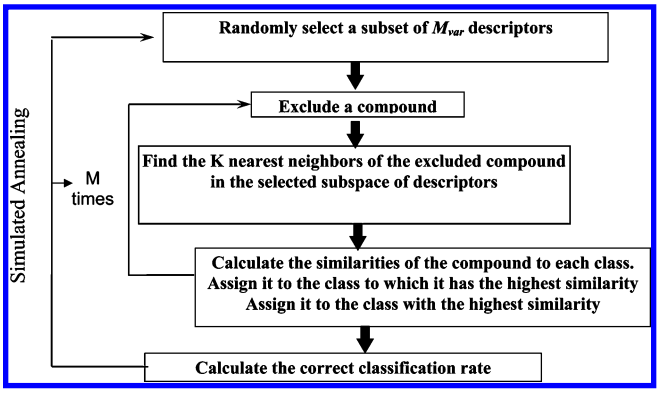

Figure 1. Flowchart of the *k*NN classification method. Simulated annealing is used to optimize the correct classification rate.

$$E_n^{\,1} = \frac{2\mathrm{TP} \times N_0}{\mathrm{TP} \times N_0 + \mathrm{FP} \times N_1} \text{ and } E_n^{\,0} = \frac{2\mathrm{TN} \times N_1}{\mathrm{TN} \times N_1 + \mathrm{FN} \times N_0} \quad (2)$$

***k*NN Classification.** The stochastic variable selection *k*NN classification method is based on the idea that assigning a compound to a class can be defined by the class membership of its nearest neighbors, taking into account weighted similarities between a compound and its nearest neighbors, as follows (Figure 1). Let N be the number of compounds in a data set. In the simplest case of binary classification, these compounds are distributed between classes a and b. Let n_a and n_b be the numbers of compounds in classes a and b, respectively, and m be the number of descriptors selected by the variable selection *k*NN classification procedure. The Tanimoto coefficient can be used as a similarity measure between two classes as follows:

$$T(a,b) = \frac{\sum_{i=1}^{m} \bar{D}_i^{\,a} \bar{D}_i^{\,b}}{\sum_{i=1}^{m} (\bar{D}_i^{\,a})^2 + \sum_{i=1}^{m} (\bar{D}_i^{\,b})^2 - \sum_{i=1}^{m} \bar{D}_i^{\,a} \bar{D}_i^{\,b}} \quad (3)$$

where $\bar{D}_i^{\,a}$ and $\bar{D}_i^{\,b}$ are average values of descriptor i for classes a and b, respectively,

$$\bar{D}_i^{\,a} = \frac{\sum_{j=1}^{n_a} D_{ij}^{\,a}}{n_a}$$

and

$$\bar{D}_i^{\,b} = \frac{\sum_{j=1}^{n_b} D_{ij}^{\,b}}{n_b}$$

and $D_{ij}^{\,a}$ is the descriptor value for compound j of class a.

Evidently, $T(a,a) = 1$. Let k be the number of nearest neighbors of compound i. Weighted similarities between each compound i and each class C (i.e., a, or b) are calculated as follows:

$$S_{i,C} = \sum_{p=1}^{k} \left\{ \frac{\exp(-\alpha d_{ip} / \sum_{p'=1}^{k} d_{ip'})}{\sum_{q=1}^{k} [\exp(-\alpha d_{iq} / \sum_{p'=1}^{k} d_{ip'})]} T(a_p, C) \right\} \quad (4)$$

where a_p in $T(a_p,C)$ is the class of compound p; α is a parameter, which in this study was set to 1; and d_{ip} is the distance between compound i and its pth nearest neighbor. In the leave-one-out cross-validation (LOO-CV) procedure, the similarity between compound i and each class C is calculated according to the following expression:

$$S'_{i,C} = \sum_{j=1}^{k} \left[\frac{\exp(-d_{ij})}{\sum_{j'=1}^{k} \exp(-d_{ij'})} S_{j,C} \right] \quad (5)$$

and compound i is assigned to the class which corresponds to the highest value of $S'_{i,C}$. The CCR for the training set (CCR_{train}) is calculated with formula 1.

For assigning an external compound (which was not included in the training set) to a class, it must have a reasonably high similarity to its nearest neighbors in the training set. The similarity threshold was defined as the maximum squared distance between a compound, for which the prediction is made, and its nearest neighbors of the training set. This squared distance can be defined as a sum of the average squared distance between the nearest neighbors within the training set and Z standard deviations of the squared distances from the average: $D_{max}^2 = \langle D_{near.neighb}^2 \rangle + Z\sigma^2_{near.neighb}$. The threshold is referred to here as the Z-cutoff.

The classification accuracy of the model is estimated using the test set as follows:

1. For each compound of the test set, k nearest neighbors from the training set are found.

2. All compounds of the test set are selected for which the distances to their nearest neighbors in the training set were within the defined Z-cutoff.

3. The similarity of each compound chosen in step 2 to each class is calculated using formula 5. The compound is assigned to a class to which it has the highest similarity.

4. The classification accuracy of the model is characterized by the CCR for the test set (CCR_{test}) calculated with formula 1.

In this study, the Z-cutoff was set equal to 6. The maximum Z-cutoff value, for which a reliable prediction of the new compounds can be obtained, is a characteristic of the applicability domain[19] of a QSAR model. As we shall see, when Z-cutoff = 6 is used, high CCR_{test} values have been obtained for several combinations of methods and descriptor types. Thus, the applicability domain for these models is characterized by at least a Z-cutoff of 6.

kNN classification QSAR is a stochastic variable selection procedure based on the simulated annealing approach. The procedure is aimed at the development of a model with the highest fitness (CCR_{train}). The procedure starts with the random selection of a predefined number of descriptors out of all descriptors. A compound excluded in the LOO-CV procedure is assigned to a class corresponding to the highest $S_{i,C}$ (see formula 4), where i is the number of the excluded compound. After each run, the cross-validated CCR_{train} is defined (see formula 1) and a predefined number of descriptors are replaced by the same number of randomly chosen descriptors from the original population. The new value of CCR_{train} is obtained using the modified subset of descriptors. If $CCR_{train}(new) > CCR_{train}(old)$, the new subset of descriptors is accepted. If $CCR_{train}(new) \leq CCR_{train}(old)$, the new subset of descriptors is accepted with the probability $p = \exp[CCR_{train}(new) - CCR_{train}(old)]/T$, and rejected with the probability $(1 - p)$, where T is the simulated annealing "temperature" parameter. During the process, T decreases until the predefined value is achieved. Thus, CCR_{train} is optimized. In the prediction process, the final set of descriptors selected is used and the expression 5 is applied. This implementation is similar to that reported for the continuous kNN QSAR method developed in our laboratory earlier.[52]

In all of the calculations reported in this work, $k_{max} = 5$, $T_{max} = 100$, $T_{min} = 10^{-9}$, the temperature scaling factor was 0.95, and the number of descriptor replacements was 3. For all of the descriptor types, the number of descriptors selected by the procedure was varied from 10 to 50 with a step of 5. For each number of descriptors selected, 10 models were built. Thus, the total number of models built for one division into training and test sets was 90.

Binary Tree Classification. We have used a binary tree classification algorithm as implemented in the MOE package.[20] In the beginning, all compounds of the training set belong to one initial or root node. The method consists of two parts: tree growing and tree pruning. Tree growing is carried out by splitting the nodes according to the rules in the form $x \leq c$ (if descriptor x is a continuous variable) or $x = c$ (if it is a categorical variable), where c is the best value for splitting the node. The goal of node splitting is to make each of the child nodes contain compounds belonging predominantly to one of the classes. In other words, the goal is to maximally reduce the sum of diversities (in terms of belonging to classes) of compounds in both child nodes in comparison to the diversity of compounds in the parent node. In the MOE DT algorithm, splitting is based on the Gini index of diversity.[53] A node cannot be split if all of the compounds in it belong to the same class or if the number of compounds in it is lower than a predefined limit.

Each leaf in the tree is assigned a class j maximally represented in this leaf. The misclassification rate in node t is calculated as $r(t) = 1 - n_j/n_t$, where n_j is the number of compounds of class j and n_t is the total number of compounds in the node. The total misclassification rate $R(T) = N_{misclass}/N_{tot}$, where $N_{misclass}$ and N_{tot} are the total number of misclassified compounds and the total number of compounds in the data set, respectively. If classes have different sizes, the misclassification rate is multiplied by weights defined as $w_j = N_{tot}/N_j$, where N_j is the number of compounds in class j.[20]

An initially grown tree can be very large and can have a very high correct classification rate for the training set.

However, usually, it performs poorly for the test set. Tree pruning is a procedure used to decrease the size of the tree and increase its classification accuracy for the test set.[20] This procedure was performed using the test set. By pruning branches of the tree, the accuracy of classification for the training set and the size of the tree are decreased. A modified tree misclassification rate $R_a(T) = R(T) + aL(T)$ is defined, where $L(T)$ is the number of leaves in the tree and $a > 0$ is a parameter.[20] According to this equation, the size of the tree and the misclassification rate are balanced. By increasing a, smaller trees can be found for which $R'_a(T) = R_a(T)$. Pruning is performed by finding a sequence of successively smaller trees T_i, starting from the initially grown tree. The smallest tree T_N is just the root node. The misclassification rate for each T_i is calculated using this test set. The output of the procedure is the tree with $R(T)$ within a specified number of standard errors of the minimum of all subtree $R(T)$ values. Standard error is defined as $\sigma = \sqrt{p(1-p)N_{\text{test}}}$, where p is the proportion of correctly classified cases. This subtree is referred to as the best subtree.[20]

The class of a new compound is predicted by the class of the leaf this compound belongs to. Classes assigned to compounds of the external test set were used to estimate the model classification accuracy. The following parameters have been used: a minimum node split size of 10, an ordered threshold of 6, and a best tree threshold of 0.5. Since the implementation of this method in the MOE package does not allow automatic variable selection, all descriptors of each type were used to build decision trees. Thus, in total, an attempt to build four decision trees, one for each type of descriptors, was undertaken.

Binary QSAR. Binary QSAR is a recent technique developed by P. Labute[21] and implemented in the MOE package.[20] This approach can be applied, if the activities y_i of compounds take only two values, zero or one, which correspond to inactive and active compounds, respectively. Binary QSAR is based on the Bayesian inference technique, which is used to classify a compound as active or inactive. Let m be the total number of compounds, and m_0 and m_1 are the numbers of inactive and active compounds, respectively ($m = m_0 + m_1$). If descriptors $X_1, X_2, ..., X_n$ are not correlated, then the conditional probability that a compound with descriptor values $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ is active, that is, $p(x) = \Pr(Y = 1|X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$, can be estimated as[20,21]

$$p(x) = \left[1 + \frac{m_0 + 1}{m_1 + 1} \prod_{j=1}^{n} \frac{f_j(x_j,0)}{f_j(x_j,1)}\right]^{-1} \quad (6)$$

where $f(x,y) = \Pr(X_j = x_j|Y = y)$. Without a loss of generality, it is assumed that descriptors $X_1, X_2, ..., X_n$ have a mean value of zero and variance 1.

Each function $f(x)$ can be estimated by considering a histogram of observed descriptor values on a set of B bins $(b_0b_1), ..., (b_{B-1}b_B)$, where $b_0 = -\infty$ and $b_B = +\infty$. The number of compounds within bin k

$$B_k = \sum_{i=1}^{m} \int_{b_{k-1}}^{b_k} \delta(x - x_i)\, dx$$

can be smoothed by approximating each δ function with a Gaussian distribution with variance σ^2:[20,21]

$$B_k = E_k - E_{k-1},\ E_k = \frac{1}{2}\sum_{i=1}^{m} erf\left(\frac{b_k - x}{\sigma\sqrt{2}}\right)$$

Finally, $f(x)$ can be estimated as[20]

$$\hat{f}(x) = \sum_{k=1}^{B} \frac{B_k + 1/c}{c + B/c}[E_k - E_{k-1}] \quad (7)$$

where

$$c = \sum_{k=1}^{B} B_k$$

In the same way, all $f_j(x_j,0)$ and $f_j(x_j,1)$ can be estimated and

$$p(x) = \left[1 + \frac{m_0 + 1}{m_1 + 1} \prod_{j=1}^{n} \frac{\hat{f}_j(x_j,0)}{\hat{f}_j(x_j,1)}\right]^{-1} \quad (8)$$

Thus, the whole binary QSAR procedure consists of the following steps:[20,21] (i) a principal component analysis of the descriptor matrix to produce a variance−covariance matrix of $x_i = Q(d_i - u)$ equal to the identity matrix, where x_i are principal components and $d_i = (d_{i1}, ..., d_{in})$ are descriptor values for compound i, and (ii) estimation of the binary QSAR model $p(x)$ parameters. The probability that a compound with descriptors d_i^{new} is active can be estimated as $p[Q(d_i^{\text{new}} - u)]$.

The binary QSAR models were built with the number of principal components varied from 5 to 40 with a step of 5 and binary threshold of 0.5. Thus, in total, eight models for each type of descriptor were built.

Support Vector Machines (SVM).[22] The SVM method tries to find the best hyperplane able to separate data in the feature space into classes. The application of SVM to the binary classification was implemented as follows. Let m be a number of points representing compounds scattered in an n-dimensional descriptor space. Compounds can be active (activity is equal to +1) or inactive (activity is equal to −1). The problem is to divide active and inactive compounds by a hyperplane in the descriptor space. If the solution of this problem is possible, the data set is referred to as separable. Otherwise it is nonseparable. If a data set is separable, a solution can be found as follows. The equation of any hyperplane in the descriptor space can be represented as $(\mathbf{wx}) - b = 0$, where $\mathbf{w}$ is a normal to the hyperplane, $\mathbf{x}$ is a vector with the beginning in the origin and end on the hyperplane, and $(\mathbf{wx})$ is the dot product of $\mathbf{w}$ and $\mathbf{x}$. Let it be the dividing hyperplane. Without a loss of generality, we can assume that for any point x_i with activity $y_i = 1$ $(\mathbf{wx}_i) + b \geq +1$, and for all points x_i with $y_i = -1$ $(\mathbf{wx}_i) + b \leq -1$. These two inequalities can be combined into one:

$$y_i[(\mathbf{wx}_i) + b] - 1 \geq 0 \quad (9)$$

The distance between the hyperplane and the closest of its

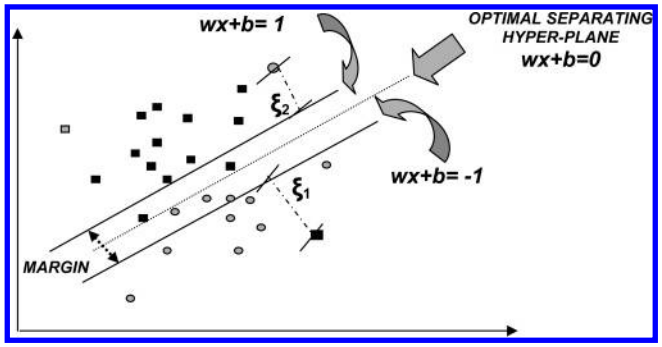

Figure 2. The separating hyperplane (**w**, *b*), limiting conditions (**wx** ± 1), and errors (ξ) for a two-dimensional training set.

data set points is equal to $1/||\mathbf{w}||$, where $||\mathbf{w}||$ is the norm of **w**. Thus, by minimizing $||\mathbf{w}||$ or $||\mathbf{w}||^2$ with the constraints in eq 9, the optimal dividing hyperplane can be found. This optimization problem can be solved by minimizing the Lagrangian

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m}\alpha_i\{y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1\} \qquad (10)$$

where $\alpha_i \geq 0\ \forall\ i$ are the Lagrange multipliers. Methods of solving this problem and finding **w** and *b* are described in ref 54. For all $y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1 > 0$, $\alpha_i = 0$, and for all $y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1 = 0$, $\alpha_i > 0$. Points for which $\alpha_i > 0$ are called support vectors. These points belong to hyperplanes $y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1 = 0$. In fact, only these points are necessary to build the optimal dividing hyperplane. Assigning compounds to a class of actives or inactives can be carried out by finding y_i from inequality 9.

In practice, if the number of points is lower than the number of descriptors minus one and no $K + 2$ points belong to a *K*-dimensional hyperplane, the data set is always separable. So, if the number of descriptors is higher than the number of compounds minus one, there is a high risk of overfitting. The hyperplane will perfectly separate points of the training set, whereas there will be poor separation of the test set. In this case, the same approach is applied as in the case when the solution does not exist, namely, such a hyperplane is sought that divides active and inactive compounds with the classification error minimized (Figure 2). The constraints in eq 9 are replaced by the inequalities

$$y_i[(\mathbf{w}\mathbf{x}_i) + b] \geq 1 - \xi_i, \qquad \xi_i \geq 0 \qquad (11)$$

where ξ_i ($i = 1, ..., m$) are slack variables. The optimal hyperplane can be found by minimizing the Lagrangian

$$L = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{m}\alpha_i\{y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1 + \xi_i\} + Cf(\sum_{i=1}^{m}\xi_i) - \sum_{i=1}^{m}\mu_i\xi_i \qquad (12)$$

where $0 \leq \alpha_i \leq C\ \forall\ i$ and μ_i are the Lagrange multipliers for the constraint $\xi_i \geq 0$. Penalty function $f(\sum_{i=1}^{m}\xi_i)$ is a positive monotonically increasing function of each parameter ξ_i. We have used the penalty function in the following form:

$$f = \begin{cases} 0, \text{ if } \sum_{i=1}^{m}\xi_i \leq \epsilon \\ \sum_{i=1}^{m}\xi_i - \epsilon, \text{ if } \sum_{i=1}^{m}\xi_i > \epsilon \end{cases} \qquad (13)$$

where ϵ is a parameter. Support vectors are defined by the condition $y_i[(\mathbf{w}\mathbf{x}_i) + b] - 1 + \xi_i = 0$.

The hyperplane parameters depend on *C* and ϵ. These parameters are also varied to optimize external prediction. Thus, the SVM procedure was run multiple times to find the optimum values of *C* and ϵ. To find models with the highest classification accuracy for both training and test sets, the calculations were carried out for all combinations of *C* and ϵ with the *C* value varied from 70 to 250 with a step of 10 and ϵ varied from 0.0 to 0.3 with a step of 0.1. Thus, for each type of descriptor, $19 \times 4 = 76$ models were built.

Model Validation by Y-Randomization. Y-randomization (randomization of the response, i.e., in our case, classes) is a widely used approach to establish the model's robustness.[55] It consists of rebuilding the models using randomized activities of the training set and the subsequent assessment of the model's statistics. It is expected that models obtained for the training set with randomized activities should have significantly lower CCR values for the training or test set than the models built using the training set with real activities. If this condition is not satisfied, real models built for this training set are not reliable and should be discarded.

A Y-randomization test was performed for those combinations of methods and descriptor types that produced models with both CCR_{train} and CCR_{test} equal to or higher than 0.7. The calculations were performed three times, with the input parameters identical to those used for building models with the classes not shuffled. Models built with randomized classes were used to predict the activities of the test set. CCR_{train} and CCR_{test} values for models built with real and randomized classes were compared with each other. Using *k*NN classification QSAR and SVM, multiple models were built. Let N_{real} and N_{rand} be the total number of models built with real and randomized classes, respectively, and n_{real} and n_{rand} be the corresponding number of models with $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7$. The fractions of models with both $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7$ are $F_{real} = n_{real}/N_{real}$ and $F_{rand} = n_{rand}/N_{rand}$, respectively. The robustness of predictive models (i.e., having both $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7$) built with real classes of the training set was defined as $R = 1 - F_{rand}/F_{real}$.[19] *R* takes values from $-\infty$ to 1. If $R \geq 0.9$, predictive models are considered reliable.

RESULTS AND DISCUSSION

***k*NN Classification Modeling.** In Table 1, we present the statistics of applying the *k*NN classification method to the P-gp data set for each descriptor type. All models built with real activities of the training set had LOO-CV $CCR_{train} \geq 0.7$. In fact, the average CCR_{train} for all types of descriptors was very high (about 0.9; Table 2). At the same time, external validation using the test set shows that the majority of models appeared to have low CCR_{test} values and, therefore, cannot

Table 1. Robustness of the *k*NN Classification Models with $CCR_{train} \geq 0.7$[a]

type of descriptors	n_{real}(train)	F_{real}(train)	CCR_{train} average	n_{rand}(train)	F_{rand}(train)	R_{train}	n_{real}(test)	F_{real}(test)	n_{rand}(test)	F_{rand}(test)	R_{test}
AP	90	1.00	0.90	210	0.78	0.22	7	0.08	0	0.00	1.00
MolconnZ	90	1.00	0.87	0	0.00	1.00	3	0.03	0	0.00	1.00
MOE	90	1.00	0.88	–	–	–	0	0.00	–	–	–
VolSurf	90	1.00	0.83	192	0.71	0.29	8	0.09	0	0.00	1.00

[a] $n_{real} = 90$ and $n_{rand} = 270$ are the numbers of models built with a real and randomized dependent variable of the training set, respectively. n_{real}(train) and n_{rand}(train) are the numbers of corresponding models with $CCR_{train} \geq 0.7$. n_{real}(test) and n_{rand}(test) are the numbers of corresponding models with $CCR_{test} \geq 0.7$. F_{real}(train) and F_{rand}(train) are the fractions of models based on real and randomized dependent variables of the training set, respectively, which have $CCR_{train} \geq 0.7$. F_{real}(test) and F_{rand}(test) are the fractions of models based on real and randomized dependent variables of the training set, respectively, which have both $CCR_{train} \geq 0.7$ and $CCR_{test} \geq 0.7$. $R_{train} = 1 - F_{rand}(train)/F_{real}(train)$. $R_{test} = 1 - F_{rand}(test)/F_{real}(test)$.

Table 2. Statistics of the Best *k*NN Classification Models[a]

		training set				test set									
type of descriptors	CCR_{train}	TP	TN	FP	FN	TP	TN	FP	FN	Se	Sp	E_n^1	E_n^0	CCR_{test}	CCR_{test} rand[b]
AP	0.89	71	57	11	5	24	15	4	8	0.75	0.79	1.56	1.52	0.77	0.58
MolconnZ	0.92	71	61	7	5	23	14	5	9	0.72	0.74	1.46	1.45	0.73	0.63
MOE	0.89	72	56	12	4	19	8	11	13	0.59	0.42	1.01	1.02	0.51	–
VolSurf	0.84	69	53	15	8	23	16	3	9	0.72	0.84	1.64	1.50	0.78	0.60

[a] AP = atom pairs, MZ = MolconnZ, VS = VolSurf. Training = 144 compounds = 76 substrates and 68 nonsubstrates. Test = 51 compounds = 32 substrates and 19 nonsubstrates. TP = true positive (substrates predicted as substrates), FP = false positives (nonsubstrates predicted as substrates), FN = false negatives (substrates predicted as nonsubstrates), TN = true negative (nonsubstrates predicted as nonsubstrates), Se = sensitivity = TP/32, Sp = specificity = TN/19, $E_n^1 = (2TP \times N_0)/(TP \times N_0 + FP \times N_1)$, $E_n^0 = (2TN \times N_1)/(TN \times N_1 + FN \times N_0)$, and CCR = correct classification rate. [b] The CCR corresponds to the best test set prediction of 270 models built with the randomized dependent variable of the training set.

Table 3. Statistics of the Best Binary QSAR Models[a]

	training set		test set									
type of descriptors	PC	CCR_{train}	TP	TN	FP	FN	Se	Sp	E_n^1	E_n^0	CCR_{test}	CCR_{test} rand[b]
AP	10	0.80	21	14	5	11	0.66	0.74	1.43	1.36	0.70	0.42
MolconnZ	10	0.86	16	15	4	16	0.50	0.79	1.41	1.22	0.64	0.43
MOE	15	0.90	19	16	3	13	0.59	0.84	1.58	1.35	0.72	0.3
VolSurf	10	0.79	23	12	7	9	0.72	0.63	1.32	1.38	0.68	0.44

[a] PC is the number of principal components. Other designations are the same as in Table 2. [b] The CCR for the best prediction of the test set for 24 models built with the randomized dependent variable of the training set.

be considered as predictive models. For example, no predictive models were obtained using MOE descriptors. Thus, these results confirm that the validation of QSAR models using an external test set is an absolutely necessary part of QSAR analysis, as was stated in several earlier publications.[18,56,57]

Results of the Y-randomization test (Table 1) show that *k*NN classification models with $CCR_{test} \geq 0.7$ are reliable. Indeed, none of the models built with randomized activities of the training set had $CCR_{test} \geq 0.7$. Thus, the robustness R_{test} for AP, MolconnZ, and VolSurf descriptors was equal to one. At the same time, the results in Table 1 indicate that, for AP and VolSurf descriptors, there were a significant number of models based on randomized activities which had $CCR_{train} \geq 0.7$. For instance, R_{train} for AP and VolSurf descriptors is lower than 0.9 (Table 1); thus, these models cannot be accepted by just considering results obtained for the training set. Again, these observations speak loudly about the necessity of test set prediction for the matter of QSAR model validation.

The best *k*NN classification models are presented in Table 2. The sensitivity (Se) and specificity (Sp) for these models demonstrate that the *k*NN models had a slightly better performance when predicting nonsubstrate compounds. Classification accuracy can also be characterized by the normalized enrichment for each class (See formula 2 and Table 2). In conclusion, predictive models were obtained using combinations of *k*NN classification with AP descriptors, MolconnZ descriptors, and VolSurf descriptors. No predictive model was obtained using *k*NN classification combined with MOE descriptors.

Decision Tree. In the case of DT, the only tree obtained was for MOE descriptors. DT was unable to generate a decision tree for any other descriptor type. As is seen from the statistics ($CCR_{train} = 0.86$, TP = 22, TN = 12, FP = 7, FN = 10, Se = 0.69, Sp = 0.63, $E_n^0 = 1.34$, $E_n^1 = 1.29$, $CCR_{test} = 0.66$), contrary to the results presented above, the tree built with MOE descriptors showed better accuracy for the prediction of substrates than for nonsubstrates. However, external validation for this tree gave $CCR_{test} = 0.66$, which was below the acceptable threshold ($CCR_{test} \geq 0.7$). Thus, no DT model was found acceptable.

Binary QSAR. Eight binary QSAR models for each descriptor type were built (see Methods section). In Table 3 we present the statistics for the best models for each descriptor type. Among all generated models, MOE descrip-

Table 4. Training and Test Set Statistics for Best Models Built with SVM[a]

				training set				test set									
type of descriptors	C	ξ	CCR_{train}	TP	TN	FP	FN	TP	TN	FP	FN	Se	Sp	E_n^1	E_n^0	CCR_{test}	CCR_{test} rand[b]
AP	70	0.1	0.94	71	65	3	5	25	16	3	7	0.78	0.84	1.66	1.59	0.81	0.59
MolconnZ	220	0.1	0.90	67	63	5	9	19	15	4	13	0.59	0.79	1.48	1.32	0.70	0.56
MOE	120	0.0	0.84	69	60	8	7	17	14	5	15	0.53	0.74	1.34	1.22	0.63	0.52
VolSurf	100	0.2	0.88	67	61	7	9	25	16	3	7	0.78	0.84	1.66	1.59	0.81	0.53

[a] C and ξ are model parameters (see the Methods section). Other designations are the same as in Table 2. [b] The CCR corresponds to the best test set prediction made among 228 randomized models.

Table 5. Robustness of SVM Classification Models with $CCR_{train} \geq 0.7$[a]'

type of descriptors	n_{real}(train)	F_{real}(train)	n_{rand}(train)	F_{rand}(train)	R_{train}	n_{real}(test)	F_{real}(test)	n_{rand}(test)	F_{rand}(test)	R_{test}
AP	76	1.00	228	1.00	0.00	25	0.33	0	0.00	1.00
MolconnZ	76	1.00	228	1.00	0.00	7	0.09	0	0.00	1.00
MOE	76	1.00	215	0.94	0.06	0	0.00	0	0.00	1.00
VolSurf	76	1.00	210	0.92	0.08	76	1.00	0	0.00	1.00

[a] $n_{real} = 76$ and $n_{rand} = 228$. The designations are the same as those in Table 1.

Table 6. Summary of Combinatorial QSAR Approach[a]

method/ descriptors	*k*NN class		decision tree		binary QSAR		SVM class	
	CCR_{train}	CCR_{test}	CCR_{train}	CCR_{test}	CCR_{train}	CCR_{test}	CCR_{train}	CCR_{test}
atom pair	**0.89**	**0.77**	–	–	0.80	0.70	**0.94**	**0.81**
MolconnZ	0.92	0.73	–	–	0.86	0.64	0.90	0.69
MOE	0.89	0.51	0.86	0.66	**0.90**	**0.72**	0.84	0.63
VolSurf	**0.84**	**0.78**	–	–	0.79	0.68	**0.88**	**0.81**

[a] CCR_{train} and CCR_{test} for the best models.

tors appeared to yield better models with the best $CCR_{test} = 0.72$. The Se and Sp values of these models indicate that, in general, binary QSAR models had higher classification accuracy for nonsubstrates than for substrates, with the exception of VolSurf descriptors, where half of the generated models predicted better for the substrate class.

The results of Y-randomization (Table 3) show that binary QSAR models are reliable since none of the models built with randomized dependent variables of the training set were able to predict the test set with $CCR_{test} \geq 0.7$.

Support Vector Machines. In Table 4 we present the statistics of the SVM classification models for each type of descriptors. All models built with real activities of the training set have LOO-CV $CCR_{train} \geq 0.7$ (Table 5). In fact, the average CCR_{train} for all types of descriptors was high (about 0.9; Table 4). At the same time, external validation using the test set shows that not all models had $CCR_{test} \geq 0.7$, even though all models generated with VolSurf descriptors had predicted the test set compounds above the acceptable threshold ($CCR_{test} \geq 0.7$). Similarly to the *k*NN method, the combination of the SVM method and MOE descriptors did not provide any acceptable model.

The results of the Y-randomization test (Table 4) show that SVM classification models with $CCR_{test} \geq 0.7$ are reliable. Indeed, none of the models built with randomized activities of the training set had $CCR_{test} \geq 0.7$. Thus, the robustness R_{test} for AP, MolconnZ, and VolSurf descriptors was equal to one (Table 5). At the same time, the results in Table 5 show that, for AP and VolSurf descriptors, there was a considerable number of models based on randomized activities which had $CCR_{train} \geq 0.7$, although none of them predicted the test set with $CCR_{test} \geq 0.7$, suggesting that these models are reliable.

The best SVM classification models are presented in Table 4. The Se and Sp for these models demonstrate that the SVM models had higher classification accuracy for P-gp nonsubstrates than for substrates. This result is also obvious from the values of the normalized enrichment for each class (cf. formula 2 and Table 4). This trend corroborates previous P-gp results published by the Penzotti[11] and Stouch[12] research groups.

In conclusion, a considerable proportion of predictive models were obtained using combinations of SVM classification with AP descriptors, and with VolSurf descriptors; the latter had an impressive performance, having all models predictive. None of the models built with the SVM method and MOE descriptors was accepted after the external test set validation.

The summary of the results of the combinatorial QSAR analysis of the P-gp data set is given in Table 6.

Comparison of the Combinatorial QSAR Models with the Pharmacophore-Based Model. A comparison between the best models obtained by the combinatorial QSAR approach and those in the previous QSAR study of the same data set[11] is given in Table 7. We can see that, in general, the acceptable combinatorial QSAR models have significantly higher classification accuracy than the pharmacophore-based model.[11] SVM models were the best: they predicted the test set with an accuracy much higher than 0.66 (which was the accuracy of the pharmacophore-based model). With the exception of the binary QSAR method in combination with VolSurf descriptors, all predictive models classified

Table 7. Comparison between Best Combinatorial QSAR Models and the Pharmacophore Model[11]

pharmacophore model[11]		combinatorial QSAR							
substrates TP	nonsubstrates TN	*k*NN VolSurf		decision tree MOE		binary QSAR MOE		SVM VolSurf	
17	15	TP	TN	TP	TN	TP	TN	TP	TN
$CCR^1 = (TP + TN)/51 =$ **0.63**[a]		23	16	22	12	19	16	25	16
$CCR_{test} = 0.5(TP/32 + TN/19) =$ **0.66**		$CCR_{test} =$ **0.78**		$CCR_{test} =$ **0.66**		$CCR_{test} =$ **0.72**		$CCR_{test} =$ **0.81**	

[a] CCR^1 is the correct classification rate for the test set as reported in ref 11.

P-gp substrates better the nonsubstrates. These results are in qualitative agreement with the Penzotti et al study of the same data set.[11]

CONCLUSIONS

We have applied a combinatorial QSAR approach to the data set of 195 P-glycoprotein substrates and nonsubstrates. The approach explores various combinations of optimization methods and descriptor types and includes rigorous and consistent model validation. The following four QSAR methods have been used: *k*NN classification QSAR, binary QSAR, decision tree, and SVM. Each method has been used with each of the following four descriptor sets: atom pair descriptors,[24,25] MolconnZ,[23] MOE,[20] and VolSurf[26] descriptors. Thus, in total, 16 (4 methods × 4 descriptor types) combinations of methods and descriptor types have been employed.

The best *k*NN classification models (Table 6) were obtained with AP descriptors (correct classification rate for the training and test sets were $CCR_{train} = 0.89$ and $CCR_{test} = 0.77$, respectively) and VolSurf descriptors ($CCR_{train} = 0.84$ and $CCR_{test} = 0.78$). The best SVM models were built with AP descriptors ($CCR_{train} = 0.94$ and $CCR_{test} = 0.81$) and VolSurf descriptors ($CCR_{train} = 0.88$ and $CCR_{test} = 0.81$). The best binary QSAR models were built with AP descriptors ($CCR_{train} = 0.80$ and $CCR_{test} = 0.70$) and MOE descriptors ($CCR_{train} = 0.90$ and $CCR_{test} = 0.72$). The decision tree method was capable of building a tree with MOE descriptors only, which appeared to have low classification accuracy for the test set ($CCR_{test} = 0.66$). Thus, no predictive models were built using the decision tree method. All predictive models had higher accuracy in the classification of P-gp nonsubstrates than substrates. This result is in agreement with the Penzotti et al.[11] and Stouch and Gudmundsson[12] models for P-gp data sets. However, our best models have significantly higher classification accuracy (0.81 for the best model) than the pharmacophore-based model of Penzotti et al.[11] built for the same data set with the reported classification accuracy of only 0.63 (or 0.66 if CCR is calculated using formula 1; cf. Table 7).

To estimate the classification accuracy of the QSAR models, we have used the CCR criterion, which takes into account that the number of compounds belonging to different classes is different. We have shown that the standard classification accuracy definition as the ratio of the number of correctly classified compounds to the total number of compounds overestimates the CCR.

We have demonstrated that by using the combinatorial QSAR approach it is possible to identify predictive models, however when only one method and one descriptor type is used there is a high chance of failure (see Table 6). The combinatorial QSAR approach allows full automation, and this project is in progress in our laboratory. Currently, only *k*NN classification and SVM QSAR are fully automated. The work is in progress to include other optimization techniques into a automated Combi-QSAR workflow.

ACKNOWLEDGMENT

We express our thanks to EduSoft for providing us with the MolconnZ program and to Chemical Computing Group for the MOE software grant. We are grateful to Dr. Gabriele Cruciani for providing us with the software to generate the VolSurf descriptors in this project, to Dr. Michele Lamb for providing us with the SMILES strings for the data set used in this study, and to Dr. Robert D. Coner for the valuable discussions. These studies were supported in part by the NIH research grant GM066940.

REFERENCES AND NOTES

(1) Sharom, F. J. The P-Glycoprotein Efflux Pump: How Does It Transport Drugs? *J. Membr. Biol.* **1997**, *160*, 161−175.

(2) Thiebaut, F.; Tsuruo, T.; Hamada, H.; Gottesman, M. M.; Pastan, I.; Willingham, M. C. Cellular Localization of the Multidrug-Resistance Gene Product P-Glycoprotein in Normal Human Tissues. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7735−7738.

(3) Eisenblatter, T.; Huwel, S.; Galla, H. J. Characterization of the Brain Multidrug Resistance Protein (BMDP/ABCG2/BCRP) Expressed at the Blood−Brain Barrier. *Brain Res.* **2003**, *971*, 221−231.

(4) Thomas, H.; Coley, H. M. Overcoming Multidrug Resistance in Cancer: An Update on the Clinical Strategy of Inhibiting P-Glycoprotein. *Cancer Control* **2003**, *10*, 159−165.

(5) Persidis, A. Cancer Multidrug Resistance. *Nat. Biotechnol.* **1999**, *1*, 94−95.

(6) Lin, J. H. Drug−Drug Interaction Mediated by Inhibition and Induction of P-glycoprotein. *Adv. Drug Delivery Rev.* **2003**, *55*, 53−81.

(7) Bodo, A.; Bakos, E.; Szeri, F.; Varadi, A.; Sarkadi, B. The Role of Multidrug Transporters in Drug Availability, Metabolism and Toxicity. *Toxicol. Lett.* **2003**, *140−141*, 133−143.

(8) 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.

(9) Masuda, S.; Uemoto, S.; Hashida, T.; Inomata, Y.; Tanaka, K.; Inui, K. Effect of Intestinal P-Glycoprotein on Daily Tacromilus through Level in a Living-Donor Small Bowel Recipient. *Clin. Pharmacol. Ther*. **2000**, *68*, 98−103.

(10) Seelig, A. A General Pattern for Substrate Recognition by P-Glycoprotein. *Eur. J. Biochem.* **1998**, *251*, 252−261.

(11) 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.

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

(13) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A Rapid Computational Method for Lead Evolution: Description and Application to α_1-Adrenergic Antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(14) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(15) Van Drie, J. H.; Nugent, R. A. Addressing the Challenges of Combinatorial Chemistry: 3D Databases, Pharmacophore Recognition and Beyond. *SAR QSAR Environ. Res.* **1998**, *9*, 1−21.
(16) Mason, J. S.; Morize, I.; Menard, I. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(17) Smellie, A.; Stanton, R.; Henne, R.; Teig, S. Conformational Analysis by Intersection: CONAN. *J. Comput. Chem.* **2003**, *24*, 10−20.
(18) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(19) Kovatcheva, A.; Golbraikh, A.; Oloff, S.; Xiao, Y.-D.; Zheng, W.; Wolschann, P.; Buchbauer, G.; Tropsha, A. Combinatorial QSAR of Ambergris Fragrance Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 582−595.
(20) Chemical Computing Group. http://www.chemcomp.com/software.htm (accessed Feb 2006).
(21) Stanford School of Medicine. http://helix-web.stanford.edu/psb99/Labute.pdf (accessed Feb 2006).
(22) Vapnik, V. N. In *The Nature of Statistical Learning Theory*; Springer: New York, 2000.
(23) http://www.edusoft-lc.com/molconn/manuals/400 (accessed Feb 2006).
(24) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(25) Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative Structure−Activity Relationship Analysis of Functionalized Amino Acid Anticonvulsant Agents Using k Nearest Neighbor and Simulated Annealing PLS Methods. *J. Med. Chem.* **2002**, *45*, 2811−2823.
(26) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular Fields in Quantitative Structure-Permeation Relationships: the VolSurf Approach. *THEOCHEM* **2000**, *503*, 17−30.
(27) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarities. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(28) Daylight Theory: Fingerprints. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed Feb 2006).
(29) Weininger, D. SMILES. 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(30) Syracuse Research Corporation − Environmental Science. http://esc.syrres.com/esc/docsmile.htm (accessed Feb 2006).
(31) http://www.tripos.com (accessed Feb 2006).
(32) Randić, M. On Characterization on Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(33) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.
(34) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure−Activity Analysis*; Wiley: New York, 1986.
(35) Kier, L. B. A Shape Index from Molecular Graphs. *Quant. Struct.-Act. Relat.* **1985**, *4*, 109−116.
(36) Kier, L. B. Inclusion of Symmetry as a Shape Attribute in Kappa-Index Analysis. *Quant. Struct-Act. Relat.* **1987**, *6*, 8−12.
(37) Hall, L. H.; Kier, L. B. Determination of Topological Equivalence in Molecular Graphs from the Topological State. *Quant. Struct.-Act. Relat.* **1990**, *9*, 115−131.
(38) Hall, L. H.; Mohney, B. K.; Kier, L. B. The Electrotopological State: An Atom Index for QSAR. *Quant. Struct.-Act. Relat*. **1991**, *10*, 43−51.
(39) 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.
(40) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: New York, 1999.
(41) Kellogg, G. E.; Kier, L. B.; Gaillard, P.; Hall, L. H. The E-State Fields. Applications to 3D QSAR. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 513−520.
(42) Kier, L. B.; Hall, L. H. A Differential Molecular Connectivity Index. *Quant. Struct.-Act. Relat.* **1991**, *10*, 134−140.
(43) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331−337.
(44) Wiener, H. J. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(45) Platt, J. R. Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *J. Chem. Phys.* **1947**, *15*, 419−420.
(46) Shannon, C.; Weaver, W. In *Mathematical Theory of Communication*; University of Illinois: Urbana, Illinois, 1949.
(47) Bonchev, D.; Mekenyan, O.; Trinajstic, N. Isomer Discrimination by Topological Information Approach. *J. Comput. Chem.* **1981**, *2*, 127−148.
(48) Balaban, A. T. Five New Topological Indices for the Branching of Tree-Like Graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.
(49) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(50) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(51) Golbraikh, A.; Shen, M.; Tropsha, A. Enrichment: A New Estimator of Classification Accuracy of QSAR Models. *Abstracts of papers of the American Chemical Society 223: 206-COMP*; American Chemical Society: Washington, DC, 2002; Part 1.
(52) Zheng, W.; Tropsha, A. A Novel Variable Selection QSAR Approach Based on the k-Nearest Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.
(53) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth International Group: Belmont, CA, 1984.
(54) Schölkopf, B.; Smola, J. A. Learning with Kernels. In *Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*; MIT Press: Cambridge, MA, 2002.
(55) Wold, S.; Eriksson, L. Statistical Validation of QSAR Results. In *Chemometrics Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309−318.
(56) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-Dimensional Quantitative Similarity−Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2564.
(57) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Modell.* **2002**, *20*, 269−276.

CI0504317