# Ligand-Based Models for the Isoform Specificity of Cytochrome P450 3A4, 2D6, and 2C9 Substrates

Lothar Terfloth,[†] Bruno Bienfait,[†] and Johann Gasteiger*[,†,‡]

Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052, Erlangen, Germany, and Molecular Networks GmbH, Henkestrasse 91, D-91052 Erlangen, Germany

Received January 12, 2007

A data set of 379 drugs and drug analogs that are metabolized by human cytochrome P450 (CYP) isoforms 3A4, 2D6, and 2C9, respectively, was studied. A series of descriptor sets directly calculable from the constitution of these drugs was systematically investigated as to their power into classifying a compound into the CYP isoform that metabolizes it. In a four-step build-up process eventually 303 different descriptor components were investigated for 146 compounds of a training set by various model building methods, such as multinomal logistic regression, decision tree, or support vector machine (SVM). Automatic variable selection algorithms were used in order to decrease the number of descriptors. A comprehensive scheme of cross-validation (CV) experiments was applied to assess the robustness and reliability of the four models developed. In addition, the predictive power of the four models presented in this paper was inspected by predicting an external validation data set with 233 compounds. The best model has a leave-one-out (LOO) cross-validated predictivity of 89% and gives 83% correct predictions for the external validation data set. For our favored model we showed the strong influence on the predictivity of the way a data set is split into a training and test data set.

## 1. INTRODUCTION

Chemoinformatic methods can provide an important contribution to drug discovery. Therefore, in many cases they have already been integrated into the drug discovery workflow. Among others, the field of in silico prediction of ADMET (absorption, distribution, metabolism, elimination, and toxicity) properties is of special interest. In silico ADMET is expected to detect and eliminate compounds with inappropriate pharmacokinetic properties at an early stage of the drug discovery process. A central step in the ADMET process is drug metabolism. Metabolic stability, drug toxicity, and drug−drug interactions have to be considered. It is still an enormous challenge to predict the metabolic fate of a drug. One reason for this is the complexity of the metabolic system responsible for the detoxification of xenobiotics in humans. A variety of enzymes is involved in the detoxification process, and some of the enzymes involved show polymorphism and have multimodal binding sites. For the oxidation reactions in phase I of metabolism, the cytochrome P450 isoforms play a pivotal role. In order to predict the first phase metabolite of a drug several selectivities have to be addressed: (1) Isoform specificity, i.e., which is the predominant isoform responsible for the metabolism of a drug. (2) Chemoselectivity; distinct metabolites are produced by different types of reactions at different rates. (3) Regioselectivity; multiple sites can be attacked for a given reaction type. (4) Stereoselectivity; different stereoisomers are produced at different rates.

Here, we investigate the isoform specificity for CYP3A4, CYP2D6, and CYP2C9 substrates. Cytochrome P450 substrates are likely to be metabolized by several isoforms. It is the aim of this study to predict the isoform responsible for the main route of metabolism. The work by Manga et al. served as a starting point for our study.[1] Yap et al. reported on a similar study on a different data set using consensus SVMs.[2]

CYP3A4 is known to metabolize structurally diverse, large lipophilic molecules. The site of metabolism is determined by the shape of the substrate and by chemical reactivity. Substrates of CYP2D6 possess a basic nitrogen 5−7 Å away from the lipophilic position of the site of metabolism. This position is in general located on or near an aromatic system. CYP2C9 has a preference for neutral or acidic molecules. The site of oxidation is 5−8 Å away from one or two H-bond donor/acceptors.

Recently, the in silico prediction of CYP-related metabolism of drugs has been reviewed.[3−6] The majority of models for the prediction of substrate specificity mentioned by Crivori and Poggesi[4] are based on small training sets and aim for quantitative structure metabolism relationships (QSMRs) based on experimental Michaelis constants $K_m$ values.

## 2. DATA SET

For our studies we used a data set which had been compiled by Manga et al.[1] from publications by Bertz et al.[7] and Tredger et al.[8] The data set comprises only drugs which are predominately metabolized by either CYP3A4, CYP2D6, or CYP2C9. Their class membership is taken to be non-

* Corresponding author phone: (49)-9131-85-26570; fax: (49)-9131-85-26566; e-mail: Gasteiger@chemie.uni-erlangen.de.
[†] Universität Erlangen-Nürnberg.
[‡] Molecular Networks GmbH.

10.1021/ci700010t CCC: $37.00 
Published on Web 07/03/2007

overlapping and is determined by the isoform responsible for the main route of metabolism. We stuck to the classification of the substrates as given by Manga et al.

**Data Set Preparation**. One hundred forty-nine compound names and SMILES strings were collected from the electronic version (PDF) of the paper.[1] Because the SMILES strings in the paper do not contain stereochemistry, we searched two structure databases (XENIA[9] and MDDR[10]) to find structures with stereochemistry. Unfortunately, for 106 compounds the structures differ between XENIA and MDDR or could not be found in both databases. Therefore, we searched also PubChem,[11] Crossfire Beilstein,[12] Scifinder,[13] and a medicinal chemistry book[14] to confirm the structures. In many instances, we found that the structures differ from one database to another, or in the case of PubChem, there is often more than one structure for a given compound name. When different structures were found, we chose the structure that occurs most frequently.

Although we started from the published data set,[1] the data set we ended up with differs significantly from the original publication:

- Seventeen compounds have a constitution different from the one given in the paper; this is the case for clarithromycin, cocaine, diltiazem, disopyramide, ergotamine, erythromycin, felodipine, rifabutin, saquinavir, tacrolimus, zidovudine, codeine, encainide, paroxetine, ibuprofen, lisuride, and zuclopenthixol for which the structures were corrected.

- Two SMILES from the published table could not be read (nicardipine and simvastatin) and were corrected.

- Stereochemistry was added for 95 compounds that have one or more stereo centers or unsymmetrical double bonds.

- We removed three compounds from the test set as they were already present in the training set (astemizole, cisapride, and ethinylestradiol).

Overall, the training data set comprises 146 compounds, thereof are 80 CYP3A4 substrates (55%), 45 CYP2D6 substrates (31%), and 21 CYP2C9 substrates (14%).

**Validation Data Set**. For external validation of the models we extracted 281 reactants from the Metabolite reaction database[15] using the following procedure. Metabolic reactions reported for the human species were exported from the Metabolite database. From this subset of reactions, reactants were extracted such that only one CYP450 isoform (CYP*) is reported in the ISOFORM field in the database. Further selection of compounds with CYP3A4, CYP2D6, and CYP2C9 gave 281 unique structures. Of these 281 structures, 12 compounds with the same structure but different stereochemistry were also removed. Without considering stereochemistry, 26 compounds were also contained in the Manga et al. data set and were therefore also removed. This yielded a validation data set of 233 compounds comprising 144 CYP3A4 substrates (62%), 69 CYP2D6 substrates (30%), and 20 CYP2C9 substrates (8%).

## 3. METHODS

A hierarchical structure representation scheme in combination with various model building methods was investigated in this paper. The first part of this section deals with the descriptors which were used for structure representation. The second part is focused on a brief overview on the model building methods.

**Descriptors.** Molecules can be represented by descriptors with different levels of sophistication. Simple molecular properties account only for the abundance of elements in a compound. Functional group counts take local topological information into account. Topological descriptors reflect the constitution of molecules. Descriptors related to the shape of molecules or the distribution of interatomic distances consider the 3D structure of the molecules. An extension of the software package ADRIANA.Code (version 2.0) was used for descriptor calculation.[16] The CACTVS toolkit (version 3.328) served for counting functional groups with SMARTS strings.[17] Additional descriptors were calculated with ChemAxon Calculator v4.0.4.[18] A compilation of all descriptors is given in Table 1.

**Structure Preprocessing.** No compounds of the training and the validation sets were ionized or contained small fragments such as counterions, solvent molecules, etc. CORINA is included within ADRIANA.Code and was used to add hydrogen atoms and to compute the 3D structures. Only a single 3D conformation was generated per structure.

**Substructure Based Descriptors Calculated by the CACTVS Toolkit**. Many 2C9 substrates possess one or more acidic functional groups, whereas many 2D6 compounds include one or more basic nitrogen functional group. The number of occurrence of a functional group can be computed by counting the number of substructure matches. For this task, we used SMARTS[19] strings. For 2D6 substrates, we generated six descriptors counting the number of different types of amines (Table 2). The number of basic nitrogen atoms ($n_{basic\_n}$) is computed by taking the sum of the number of aliphatic amines and the number of guanidines. Aniline and pyridine were excluded from the basic nitrogen counts because they are 4−5 orders of magnitude less basic than an aliphatic amine. The data set contains two guanidine derivatives (#124, debrisoquine, #135 phenformin); both are 2D6 substrates.

To account for the presence of acidic functional groups, we counted the occurrences of three different substructures: (1) carboxylic acid, (2) 1,3-diketone in the tautomeric enol form, and (3) acyl sulfonamides. The SMARTS patterns which were used to determine the substructure based descriptors with CACTVS are listed in Table 2.

**Descriptors Calculated by ADRIANA.Code**. Topological autocorrelation transforms the information from the structure diagram of a molecule into a vector with a fixed number of components. The component $a(d)$ of the autocorrelation vector for the topological distance $d$ results from a double summation of all atom pairs as given by eq 1:

$$a(d) = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} p_j p_i \delta(d_{t,ij}, d) \qquad \delta = \begin{cases} 1 \;\forall\; d_{t,ij} = d \\ 0 \;\forall\; d_{t,ij} \neq d \end{cases} \tag{1}$$

Here, $d_{t,ij}$ is the topological distance between atoms $i$ and $j$ (i.e., the number of bonds for the shortest path in the structure diagram), $N$ is the number of atoms in the molecule, and $p_i$ and $p_j$ are properties of atoms $i$ and $j$, respectively.

We denote the *Euclidean* distance between atom $i$ and $j$ *as* $d_{ij}$. Let the eccentricity of atom $i$, $E_i$, be the largest value of $d_{ij}$, i.e.

$$E_i = \max\{d_{ij}, 1 \le j \le N\} \tag{2}$$

**Table 1.** List of Descriptors Used in the Study

| no. | name | description | ref |
|---|---|---|---|
| | | I. Molecular Properties (A) Calculated with ADRIANA.Code | |
| 1 | MW | molecular weight | |
| 2 | HDon | number of hydrogen bonding donors derived from the sum of N−H and O−H groups in the molecule | 20 |
| 3 | HAcc | number of hydrogen bonding acceptors derived from the sum of nitrogen and oxygen atoms in the molecule | 20 |
| 4 | TPSA | topological polar surface area | 21 |
| 5 | ASA | aproximate surface area | 22 |
| 6 | $\alpha$ | mean molecular polarizability | 23−26 |
| 7 | $\mu$ | molecular dipole moment | 27 |
| 8,9 | $\chi_0, \chi_1$ | connectivity $\chi$ indices | 28 |
| 10−12 | $\kappa_1,\kappa_2,\kappa_3$ | $\kappa$ shape indices | 28 |
| 13 | W | Wiener path number | 29 |
| 14 | $\chi_R$ | Randic index | 27 |
| 15 | $D_3$ | diameter | 30 |
| 16 | $R_3$ | radius | 30 |
| 17 | $I_3$ | geometric shape coefficient | 30,31 |
| 18−20 | $\lambda_1,\lambda_2,\lambda_3$ | principal moments of inertia | 27 |
| 21 | $r_{gyr}$ | radius of gyration | 32,33 |
| 22 | $r_{span}$ | radius of the smallest sphere, centered at the center of mass which completely encloses all atoms in the molecule | 33 |
| 23 | $\epsilon$ | molecular eccentricity | 27 |
| 24 | $\Omega$ | molecular asphericity | 27 |
| 25 | $r_2$ | radius perpendicular to $D_3$ | |
| 26 | $r_3$ | radius perpendicular to $D_3$ and $r_2$ | |
| | | I. Molecular Properties (B) Calculated with CACTVS | |
| 27 | $n_{aliph\_amino}$ | number of aliphatic amino groups | |
| 28 | $n_{prim\_amino}$ | number of primary aliphatic amino groups | |
| 29 | $n_{sec\_amino}$ | number of secondary aliphatic amino groups | |
| 30 | $n_{tert\_amino}$ | number of tertiary aliphatic amino groups | |
| 31 | $n_{guanidine}$ | number guanidine groups | |
| 32 | $n_{basic\_n}$ | number of basic, nitrogen containing functional groups | |
| 33 | $n_{acidic\_groups}$ | number of acidic functional groups | |
| 34 | $n_{sulfonamide}$ | number of sulfonamide-C═O groups | |
| 35 | $n_{prim\_sec\_amino}$ | $n_{prim\_amino} + n_{sec\_amino}$ | |
| 36 | $n_{aro\_amino}$ | number of aromatic amino groups | |
| 37 | E_TPSA | topological polar surface area | |
| | | I. Molecular Properties (C) Calculated with ChemAxon | |
| 38 | $n_{acceptor}$ | number of hydrogen bond acceptors | |
| 39 | $n_{acceptor\_sites}$ | number of hydrogen bond acceptor sites | |
| 40 | $n_{aliphatic\_atoms}$ | number of aliphatic atoms | |
| 41 | $n_{aliphatic\_bonds}$ | number of aliphatic bonds | |
| 42 | $n_{aliphatic\_rings}$ | number of aliphatic rings | |
| 43 | $n_{aromatic\_atoms}$ | number of aromatic atoms | |
| 44 | $n_{aromatic\_bonds}$ | number of aromatic bonds | |
| 45 | $n_{aromatic\_rings}$ | number of aromatic rings | |
| 46 | $n_{asym\_atoms}$ | number of asymmetric atoms | |
| 47 | $n_{atoms}$ | number of atoms in the molecule | |
| 48 | $J$ | Balaban distance connectivity | |
| 49 | $n_{bonds}$ | number of bonds in the molecule | |
| 50 | $n_{aro\_carbon\_ring}$ | number of aromatic rings in the molecule containing carbon atoms only | |
| 51 | $n_{ring\_carbon}$ | number of those rings in the molecule, which contain carbon atoms only | |
| 52 | $n_{chain\_atoms}$ | number of chain atoms (non-ring atoms excluding hydrogens) | |
| 53 | $n_{chain\_bonds}$ | number of chain bonds (non-ring bonds excluding bonds of hydrogen atoms) | |
| 54 | $C$ | cyclomatic number, i.e., the number of nonoverlapping cycles | |
| 55 | $n_{donors}$ | number of hydrogen bond donors | |
| 56 | $n_{donor\_sites}$ | number of hydrogen bond donor sites | |
| 57 | $n_{fused\_ali\_rings}$ | number of aliphatic rings having common bonds with other rings | |
| 58 | H | Harary index | |
| 59 | $n_{hetero\_aro\_rings}$ | number of aromatic rings containing hetero atoms | |
| 60 | $n_{hetero\_rings}$ | number of rings containing hetero atoms | |
| 61 | WW | Hyper Wiener index | |
| 62 | $max_{ring\_size}$ | size of the largest ring in the molecule | |
| 63 | log$P$ | *n*-octanol/water partition coefficient | 34 |
| 64 | $\alpha$ | molecular polarizability | |
| 65 | PSA | polar surface area | |
| 66 | $\chi_R$ | Randic index | |
| 67 | MR | molar refractivity | 34 |
| 68 | $n_{ring\_atoms}$ | number of atoms in rings | |
| 69 | $n_{ring\_bonds}$ | number of bonds in rings | |
| 70 | $n_{rotatable\_bonds}$ | number of rotatable bonds | |
| 71 | $min_{ring\_size}$ | size of the smallest ring in the molecule | |
| 72 | SZD | Szeged index | 27 |

**Table 1** (Continued)

| no. | name | description | ref |
|---|---|---|---|
| | | I. Molecular Properties (C) Calculated with ChemAxon (Continued) | |
| 73 | $n_{tautomers}$ | number of all tautomers of a molecule | |
| 74 | $W$ | Wiener index | 29 |
| 75 | $p$ | Wiener polarity | 29 |
| 76 | p$K_a$ | acidity | |
| 77 | p$K_b$ | basicity | |
| 78−87 | log$D$ | $n$-octanol/water distribution coefficient (pH dependent); ten values for pH 1.0−10.0 | 35 |
| | | II. Vectorial Descriptors Calculated with ADRIANA.Code | |
| 88−98 | 2D-AC$_{identity}$ | topological autocorrelation; property: identity | |
| 99−109 | 2D-AC$_{\chi LP}$ | topological autocorrelation; property: lone pair electronegativity $\chi_{LP}$ | |
| 110−120 | 2D-AC$_{\chi\sigma}$ | topological autocorrelation; property: $\sigma$-electronegativity $\chi_\sigma$ | |
| 121−131 | 2D-AC$_{\chi\pi}$ | topological autocorrelation; property: $\pi$-electronegativity $\chi_\pi$ | |
| 132−142 | 2D-AC$_{q\sigma}$ | topological autocorrelation; property: $\sigma$-charge $q_\sigma$ | |
| 143−153 | 2D-AC$_{q\pi}$ | topological autocorrelation; property: $\pi$-charge $q_\pi$ | |
| 154−164 | 2D-AC$_{qtot}$ | topological autocorrelation; property: total charge $q_{tot}$ | |
| 165−175 | 2D-AC$_\alpha$ | topological autocorrelation; property: polarizability $\alpha$ | |
| 176−303 | 3D-AC | spatial autocorrelation using identity as atom property | |

**Table 2.** Definition of the SMARTS Patterns Used To Determine the Substructure Based Descriptors

| no. | name | SMARTS |
|---|---|---|
| 27 | $n_{aliph\_amino}$ | [N]([H,C&X4])([H,C&X4])([H,C&X4]) |
| 28 | $n_{prim\_amino}$ | [NH2]([H,C&X4]) |
| 29 | $n_{sec\_amino}$ | [NH]([C&X4])([C&X4] |
| 30 | $n_{tert\_amino}$ | [N]([C&X4])([C&X4])([C&X4]) |
| 31 | $n_{guanidine}$ | [N]([H,C&X4])([H,C&X4])[C&A](=[N&A][H,C&X4])[N] |
| 32 | $n_{basic\_n}$ | [N]([H,C&X4])([H,C&X4])([H,C&X4])<br>[N]([H,C&X4])([H,C&X4])[C&A](=[N&A][H,C&X4])[N] |
| 33 | $n_{acidic\_groups}$ | [OH][C,N,S,P](=O)<br>[OH][A]=[A][C,N,S,P](=O)<br>[NH]([S](=O)=O)C=O |
| 34 | $n_{acyl\ sulfonamides}$ | [NH]([S](=O)=O)C=O |
| 36 | $n_{aro\_amino}$ | [N]([a])([H,a,C&X4])([H,a,C&X4]) |

where $N$ is the number of atoms in the molecule. The geometric diameter, $D_3$, is defined to be

$$D_3 = \max\{E_i, 1 \le i \le N\} \tag{3}$$

and the geometric radius, $R_3$, is given by

$$R_3 = \min\{E_i, 1 \le i \le N\} \tag{4}$$

Petitjean[30] introduced the geometric shape coefficient $I_3$ to be

$$I_3 = (D_3 - R_3)/R_3 \tag{5}$$

In analogy to the eccentricity of planar ellipses the molecular eccentricity can be expressed in terms of the principal moments of inertia $\lambda_1$, $\lambda_2$, and $\lambda_3$. The molecular eccentricity $\epsilon$ is defined as

$$\epsilon = \frac{\{(\max[\lambda_i])^2 - (\min[\lambda_i])^2\}^{1/2}}{\max[\lambda_i]} \quad i = 1,2,3 \tag{6}$$

where $\max[\lambda_i]$ and $\min[\lambda_i]$ are the maximum and minimum eigenvalues of the inertia tensor $\boldsymbol{I}$. Spherical molecules with threefold degenerated principal moments of inertia have a molecular eccentricity of $\epsilon = 0$. For linear molecules one moment of inertia vanishes and the molecular eccentricity becomes $\epsilon = 1$.

The molecular asphericity $\Omega$ is calculated from the principal moments of inertia $\lambda_1$, $\lambda_2$, and $\lambda_3$ by the equation

$$\Omega = \frac{1}{2}\frac{\sum_{i=1}^{2}\sum_{j=i+1}^{3}(\lambda_i - \lambda_j)^2}{(\sum_{i=1}^{3}\lambda_i)^2} \tag{7}$$

Let atoms $m$ and $n$ have the largest *Euclidean* distance, $D_3$, in the molecule. The radius $r_2$ is the maximum distance between the line defined by atoms $m$ and $n$ to one of the remaining atoms in this molecule. This atom is referred to as atom $o$. The radius $r_3$ is the maximum distance between the plane defined by atoms $m$, $n$, and $o$ to one of the remaining atoms in this molecule.

3D autocorrelation can be calculated in a manner similar to topological autocorrelation. The spatial distance is used instead of the topological distance as given in eq 8:

$$a(d_l, d_u) = \frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N} p_j p_i \delta(d_{ij}, d_l, d_u)$$

$$\delta = \begin{cases} 1 \ \forall\ d_l < d_{ij} \le d_u \\ 0 \ \forall\ d_{ij} \le d \vee d_{ij} > d_u \end{cases} \tag{8}$$

The component of the autocorrelation vector $a(d_l, d_u)$ for distances in the range between the lower and upper distance boundary $d_l$ and $d_u$, respectively, is the sum of the products of an atomic property $p$ for atoms $i$ and $j$. Atoms having a Euclidian distance $d_{ij}$ within the considered interval $[d_l, d_u]$ contribute to this sum.

**Descriptors Calculated by the ChemAxon Calculator.** ChemAxon calculator v 4.0.4 was used to calculate various molecular properties.[18] Of particular interest is the calculation of p$K_a$ and p$K_b$ values. The ChemAxon calculator does not provide p$K_a$ and p$K_b$ for compounds that do not have ionizable atoms. When the p$K_a$ value was not provided, we set its value arbitrarily to 20. When the p$K_b$ value was not provided, we set its value arbitrarily to −10.

The ChemAxon calculator predicted a very acidic p$K_a$ value for three benzodiazepines derivatives (alprazolam: −13.55, midazolam: −1.35, triazolam: −13.73). These

values are obviously not correct and were replaced by the arbitrary value 20.

log*D* values were calculated for ten pH values ranging from 1.0 to 10.0.

Due to different implementations the descriptors Randic index, Wiener index, topological polar surface area, mean molecular polarizability, and number of hydrogen bond donors and acceptors, respectively, occur more than once in Table 1.

**Modeling Methods.** The data mining software Weka[36,37] provides many modeling methods, some of them have parameters. Due to time constraints, it was not possible to exhaustively test every combination of descriptor set, data modeling, and variable selection methods. We explored a subset of the Weka modeling space selecting methods that were deemed to have a good modeling power.

The modeling methods that were used in this paper are briefly described in the following:

- K-nearest neighbors search (kNN): kNN is an instance based classifier[38] that predicts the class of a test sample by giving it the same class as its nearest neighbor, that is, the most similar instance found in the learning set. K is equal to 1 or can be a larger number when looking for a set of k nearest neighbors. The metric used to search the nearest neighbors is the Euclidian distance. In Weka, variables are automatically scaled into the range [0, 1].

- C4.5/J48: J48 is a decision tree based on the C4.5 algorithm of Quinlan.[39]

- Multilayer perceptron:[40] neural network trained by the back-propagation of error algorithm. The network is built automatically using one hidden layer for which the number of neurons equals the sum of the number of descriptors and the number of classes divided by two.

- Radial basis function neural network (RBF):[41] RBF is a two-layer neural network—one hidden and one output layer. In Weka, the hidden layer contains neurons whose weights result from a K-means clustering of the training samples. Output of the hidden neurons is calculated by applying a Gaussian function on the distance between the weights of the hidden neuron and the input vector. The output layer is the same as the multilayer perceptron: the linear combination of the outputs of the hidden layer is transformed by a sigmoid function.

- Logistic regression: Logistic regression is a regression method that uses the logit transform to convert class probabilities into a regression. Simple logistic regression builds logistic regression using the LogitBoost fitting algorithm, which leads to automatic variable selection.[42]

- A support vector machine (SVM) is a linear binary classifier which selects a small number of input vectors close to the class boundary (support vectors) and seeks for an optimal hyperplane that separates the support vectors. The implementation in Weka is based on the algorithm published by Platt.[43,44]

**Cross-Validation (CV) Technique.** All models we built were checked by extensive CV. We performed LOO, 10-, 5-, 3- and 2-fold stratified CV, and finally, we used an external validation data set that is larger than the training set.

In *n*-fold CV, the data set is divided randomly into *n* partitions of similar size. One partition (or fold) is hold for testing, whereas the other partitions are used for training the model. The model is rebuilt *n* times, once for each fold ensuring that all compounds are used for testing once. Automatic variable selection is redone each time if required by the chosen modeling method. Stratification means that the data set is divided randomly such that the class distribution is similar in the training and test sets. The partition is performed with the help of a random number generator and was repeated at least ten times. This allows us to collect the average, standard deviation, and minimum and maximum of the prediction rates. We used the software package Weka to perform the LOO and *n*-fold CV.[36,37]

**Automatic Variable Selection**. We used the algorithms implemented in Weka to automatically select subsets of variables that are highly correlated with the class while having low intercorrelation (attribute evaluator CfsSubsetEval combined with either the BestFirst or the ExhaustiveSearch search method). The selection of variables was repeated for each fold during CV. The data analysis method Simple Logistic Regression and C4.5/J48 perform automatic variable selection during model building.

**Clustering**. To cluster the training and validation sets, we used a structural descriptor based on "fingerprints", string of 1 or 0 bits that account for the presence or absence of a substructure. The similarity between two compounds represented as a fingerprint is calculated by using the Tanimoto coefficient. Fingerprints (252 bits string) were calculated with the MOSES chemoinformatics toolkit.[45] We used the program SUBSET[46,47] for clustering.

## 4. RESULTS

In this paper, the following conventions for the naming of the training, test, and validation set were chosen:

train_146: includes all compounds from Table 1 from the paper of Manga et al. but with three duplicate structures removed.[1]

train_96: includes all compounds used for training by Manga et al.[1]

test_50: includes all compounds used for testing by Manga et al.[1] but with three duplicate structures removed. The prediction is performed using the model built on the train_96 set.

valid_233: validation set that includes 233 compounds that we extracted from the Metabolite database.

In this investigation a systematic build-up process for representing chemical structures was used. First, only interatomic distances of the 3D molecular models were used (model 1). Then, some shape and size descriptors as well as counts of acid and basic groups were added (model 2). This was then followed by adding a series of descriptors based on eight different electronic properties which were combined into descriptors by taking into account the number of bonds between all combinations of atoms in a molecule (model 3). We then added descriptors that counted various types of atoms and bonds as well as comprised some topological indices and included values for p$K_a$, p$K_b$, and log*D* (model 4). At each stage of structure representation a variety of data analysis methods was utilized to build predictive models. For space reasons, only the best model is reproduced here.

**Model 1: Using a Vectorial Descriptor Reflecting the Distribution of Interatomic Distances**. The first model is

**Table 3.** Predictivity of Model 1 Using Four Descriptors for the Training Set Train_146

| partition and CV | | no. of runs | % correct predictions mean | stdev | min | max |
|---|---|---|---|---|---|---|
| train_146 | | 5 | 74.7 | 0.0 | 74.7 | 74.7 |
| train_146 | LOO | 5 | 72.6 | 0.0 | 72.6 | 72.6 |
| | 10-fold | 10 | 71.6 | 1.5 | 69.1 | 73.6 |
| | 5-fold | 20 | 71.0 | 1.3 | 69.1 | 73.3 |
| | 3-fold | 33 | 69.2 | 3.6 | 61.7 | 74.7 |
| | 2-fold | 50 | 67.9 | 3.6 | 58.9 | 74.7 |
| train_96 | | 5 | 80.2 | 0.0 | 80.2. | 80.2 |
| test_50 | | 5 | 56.0 | 0.0 | 56.0 | 56.0 |

**Table 4.** Predictivity of Model 2 Established with a Decision Tree Using Three Descriptors

| partition and CV | | no. of runs | % correct predictions mean | stdev | min | max |
|---|---|---|---|---|---|---|
| train_146 | | 5 | 88.4 | 0.0 | 88.4 | 88.4 |
| train_146 | LOO | 5 | 87.7 | 0.0 | 87.7 | 87.7 |
| | 10-fold | 10 | 84.7 | 2.2 | 81.6 | 88.5 |
| | 5-fold | 20 | 83.4 | 1.9 | 80.8 | 87.7 |
| | 3-fold | 33 | 83.4 | 1.7 | 80.1 | 87.0 |
| | 2-fold | 50 | 81.6 | 3.7 | 70.5 | 86.3 |
| train_96 | | 5 | 92.7 | 0.0 | 92.7 | 92.7 |
| test_50 | | 5 | 72.0 | 0.0 | 72.0 | 72.0 |

based on descriptors reflecting the distribution of interatomic distances within the 3D models of the molecules. 3D autocorrelation vectors as given in eq 8 are suited for this task if the property identity, i.e., $p_i = p_j = 1$, is used. Interatomic distances in the range between 1.0 and 13.8 Å were sampled for the property identity with a resolution of 0.1 Å. Thus, 128 dimensional descriptor vectors served as input for model building (descriptors 176−303 of Table 1). Automatic selection of variables (exhaustive search) combined with multinomal logistic regression as implemented in Weka was used as model building method.

For the training data set (train_146), four descriptors resulted from the variable selection procedure. The 3D autocorrelation vector components 1, 2, 3, and 48 representing the distance intervals [1.0−1.1] Å (C−H bonds), [1.1−1.2] Å (this distance is only observed in clarithromycin, cyclosporin, erythromycin, and ethinylestradiol), [1.2−1.3] Å (e.g., N═N, C═O), and [5.8−5.9] Å were selected. In the LOO CV, components 1, 3, and 48 were always selected. Component 2 was selected 136 times or 93% of all cases. In addition, component 26 was part of the subset six times, and component 41 was selected once.

As can be seen from Table 3 the training data set (train_146) is predicted with an accuracy of 74.7%. Depending on the fold size the predictivity drops down only by 2−7% for the CV runs. Nevertheless, the prediction on the test data set test_50 is only 56.0%. This drop in predictivity for the published test data set will be addressed in the discussion.

**Model 2: Using a Combination of 3D Autocorrelation, Global Molecular, Shape/Size-Related Descriptors, and Substructure Counts.** The descriptor set from model 1 was extended by additional global molecular descriptors which reflect the compounds shape and size as well as substructure-based descriptors characterizing a compounds acidity or basicity (descriptors 4, 5, 7, 15−37, and 176−303 from Table 1.

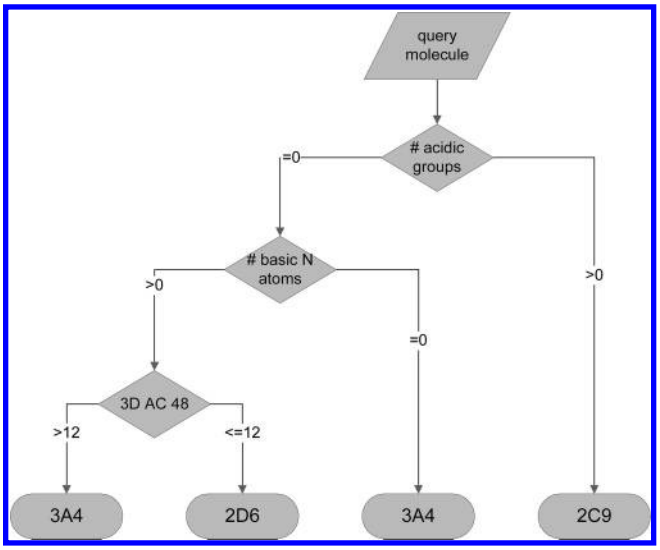



**Figure 1.** Decision tree for model 2. 3D AC 48 is the 3D autocorrelation vector component for the distance interval [5.8−5.9] Å.

An exhaustive search for the best subset of the 154 descriptors resulted in a subset of six descriptors: number of basic nitrogen atoms (descriptor 32), number of acidic groups (descriptor 33), number of secondary aliphatic amino groups (descriptor 29), and 3D autocorrelation vector components 3, 16, and 48.

The best model was built by combining the automatic variable selection (exhaustive search) with a C4.5/J48 decision tree for which the minimum number of instances per leaf was set to 6. The tree obtained needs only three descriptors for the entire data set (Figure 1).

Inclusion of the shape, size, and substructure-based descriptors improved the predictivity of this model remarkably in comparison to model 1. It is noteworthy that the predictivity is 88.4% for the training run even though only three descriptors occur in the decision tree. The average predictivity of the model decreases to 81.6% for the 2-fold CV.

**Model 3: Using a Combination of Topological Autocorrelation, 3D Autocorrelation, Global Molecular, Shape/Size-Related Descriptors, and Substructure Counts.** The descriptor set used within this model is an extension of model 2 and comprises all vectorial descriptors (topological and 3D autocorrelation), substructure-based descriptors, and a manual selection of shape and size related descriptors (descriptors: 4, 5, 7, 15−37, and 88−303 from Table 1). Topological autocorrelation vectors were calculated for the topological distances 0−10 and eight physicochemical atomic properties: identity, lone pair electronegativity $\chi_{LP}$, $\sigma$-electronegativity $\chi_\sigma$, $\pi$-electronegativity $\chi_\pi$, $\sigma$-charge $q_\sigma$, $\pi$-charge $q_\pi$, total charge $q_{tot}$, and polarizability $\alpha$. The total number of descriptor components is 242. Automatic variable selection using the best first method yielded a subset of 12 descriptors for the training set (Table 5). The occurrence of how often these descriptors were chosen in the LOO CV run is given in the last column of this table. Unlike the two previous models, we did not use the exhaustive search for variable selection because of the computation time required for a larger number of descriptors.

The model that gave the best results combined automatic variable selection (best first search) with a SVM for which

**Table 5.** Descriptors Selected by Automatic Variable Selection for the Training Set in Model 3 and Percentage of Their Occurrence in a LOO Cross-Validated Variable Selection

| descriptor no. (Table 1) | name | description | | | occurrence in LOO [%] |
|---|---|---|---|---|---|
| 94 | 2D-$AC_{identity}$(5) | top. autocorrelation | identity | top. distance 5 | 98 |
| 145 | 2D-$AC_{q\pi}$(3) | top. autocorrelation | $\pi$-charge | top. distance 3 | 100 |
| 148 | 2D-$AC_{q\pi}$(6) | top. autocorrelation | $\pi$-charge | top. distance 6 | 100 |
| 126 | 2D-$AC_{\chi\pi}$(5) | top. autocorrelation | $\pi$-electronegativity | top. distance 5 | 97 |
| 133 | 2D-$AC_{q\sigma}$(1) | top. autocorrelation | $\sigma$-charge | top. distance 1 | 58 |
| 134 | 2D-$AC_{q\sigma}$(2) | top. autocorrelation | $\sigma$-charge | top. distance 2 | 100 |
| 116 | 2D-$AC_{\chi\sigma}$(6) | top. autocorrelation | $\sigma$-electronegativity | top. distance 6 | 100 |
| 223 | 3D-$AC_{identity}$([5.8−5.9] Å) | spatial autocorrelation | identity | [5.8−5.9] Å | 79 |
| 33 | $n_{acid\_groups}$ | number of acidic groups | | | 100 |
| 27 | $n_{aliph\_amino}$ | number of aliphatic amino groups | | | 79 |
| 32 | $n_{basic\_n}$ | number of basic, nitrogen containing functional groups | | | 100 |
| 26 | $r_3$ | radius | | | 76 |

**Table 6.** Predictivity of Model 3 Using 12 Descriptors for the Training Set Train_146

| partition and CV | | no. of runs | % correct predictions mean | stdev | min | max |
|---|---|---|---|---|---|---|
| train_146 | | 5 | 90.4 | 0.0 | 90.4 | 90.4 |
| train_146 | LOO | 5 | 89.0 | 0.0 | 89.0 | 89.0 |
| | 10-fold | 10 | 87.8 | 2.3 | 83.4 | 90.4 |
| | 5-fold | 20 | 87.8 | 1.8 | 83.6 | 90.5 |
| | 3-fold | 33 | 86.3 | 1.9 | 82.2 | 90.4 |
| | 2-fold | 50 | 84.2 | 3.2 | 73.3 | 91.1 |
| train_96 | | 5 | 90.6 | 0.0 | 90.6 | 90.6 |
| test_50 | | 5 | 82.0 | 0.0 | 82.0 | 82.0 |

the exponent of the polynomial kernel was set to two. The results are shown in Table 6. In comparison to model 2 the inclusion of the topological autocorrelation vectors slightly improves the predictivity of the model. In the LOO CV 89.0% of all compounds are correctly classified. As expected, the smaller the number of folds in the CV experiments the smaller is the predictivity of the model. The minimum predictivity is observed for the 2-fold CV with an average predictivity of 84.2% in 50 repetitions.

**Model 4: Using All Descriptors**. The last model is an extension of model 3. It was built using all 303 descriptors of Table 1. The best CV results were obtained using the combination of automatic variable selection (best first search) with a SVM for which the exponent of the polynomial kernel was set to 2. For the training set train_146, the automatic variable selection yielded a set of 15 descriptors (Table 7) very similar to model 3 (Table 5). For this model 4, three ChemAxon descriptors were added: p$K_a$, p$K_b$, and log$D$ calculated for a pH equal to 1.0. The number of aliphatic amino groups used in model 3 has been replaced by the number of secondary aliphatic amines.

Prediction accuracies are shown in Table 8. In comparison to model 3 the inclusion of the ChemAxon descriptors only slightly improves the predictivity of the model. In the LOO CV 90.4% of all compounds are correctly classified. As seen previously, the minimum predictivity obtained for the 2-fold CV is lower than for LOO and has an average value of 84.5% for 50 repetitions.

**Diversity and Overlap of the Training and Validation Set.** To estimate the relative diversity and overlap of the training and validation sets, we clustered both sets and counted the number of clusters. This approach can provide a simple way for estimating the diversity of a set of compounds.[48] The number of clusters obtained using a Tanimoto coefficient of 80% were 91 and 120 for the training and validation set, respectively. The larger number of clusters for the validation set indicates that the latter is more diverse than the former. Clustering several compounds sets together with a given similarity threshold can be used to deduce the structural overlap of the compound libraries.[49] We clustered the combined training − validation set, and the number of clusters obtained was 172, which is smaller than the sum of number of clusters of the two sets. The difference between (91 + 120) and 172 (39) gives an estimation of the overlap of the two data sets (Figure 2).

With 233 compounds, the validation set is about 60% larger than the training set. The clustering study indicates that the validation set is also more diverse than the training set.

**Validation of the Models with an External Data Set Extracted from the Metabolite Database.** The isoform responsible for metabolism was predicted for the 233 compounds of the validation set for each model (Table 9).

Model 1 resulted in 68% correct prediction for the validation set. Models 2−4 raised the prediction rates to 80−83%, which are slightly lower than the 2-fold CV results. Taking into account that the external validation set is more diverse and larger than the training set, the prediction accuracy for the validation set is higher as one might expect.

LOO CV of the validation set using the modeling method of model 3 gave 80.7% correct predictions which can be regarded as a boundary of the prediction accuracy to be expected for the external validation set. The external validation of model 3 results in an even higher percentage of correct predictions of 82.9%.

The external validation of our models demonstrates that they can predict isoform specificity for compounds that are structurally different from those in the training set. The compounds in the validation set taken from the metabolite database are within the scope of models 1−4.

## 5. DISCUSSION

**Model Comparison.** Model 1 is based on a simple, vectorial descriptor and allows an accuracy of predictions of 78.0% in a LOO CV.

Model 2 increases the predictivity to 88.4% and consists of a very simple decision tree having only six branches and four leaves for the training set. Initially, six descriptors out of 154 were automatically selected by the Weka automatic

**Table 7.** Descriptors Selected by Automatic Variable Selection for the Training Set in Model 4 and Percentage of Their Occurrence in a LOO Cross-Validated Variable Selection

| descriptor no. (Table 1) | name | description | | | occurrence in LOO [%] |
|---|---|---|---|---|---|
| 94 | 2D-$AC_{identity}(5)$ | top. autocorrelation | identity | top. distance 5 | 97 |
| 145 | 2D-$AC_{q\pi}(3)$ | top. autocorrelation | $\pi$-charge | top. distance 3 | 100 |
| 148 | 2D-$AC_{q\pi}(6)$ | top. autocorrelation | $\pi$-charge | top. distance 6 | 100 |
| 126 | 2D-$AC_{\chi\pi}(5)$ | top. autocorrelation | $\pi$-electronegativity | top. distance 5 | 95 |
| 133 | 2D-$AC_{q\sigma}(1)$ | top. autocorrelation | $\sigma$-charge | top. distance 1 | 75 |
| 134 | 2D-$AC_{q\sigma}(2)$ | top. autocorrelation | $\sigma$-charge | top. distance 2 | 100 |
| 116 | 2D-$AC_{\chi\sigma}(6)$ | top. autocorrelation | $\sigma$-electronegativity | top. distance 6 | 97 |
| 223 | 3D-$AC_{identity}$([5.8−5.9] Å) | spatial autocorrelation | identity | [5.8−5.9] Å | 79 |
| 33 | $n_{acid\_groups}$ | number of acidic groups | | | 100 |
| 29 | $n_{sec\_amino}$ | number of aliphatic amino groups | | | 97 |
| 32 | $n_{basic\_n}$ | number of basic, nitrogen containing functional groups | | | 100 |
| 26 | $r_3$ | radius | | | 99 |
| 76 | p$K_a$ | acidity | | | 100 |
| 77 | p$K_b$ | basicity | | | 100 |
| 78 | log*D* pH = 1.0 | n-octanol/water distribution coefficient at pH=1 | | | 100 |

**Table 8.** Predictivity for Model 4 Using 15 Descriptors for the Training Set Train_146

| partition and CV | | no. of runs | % correct predictions mean | stdev | min | max |
|---|---|---|---|---|---|---|
| train_146 | | 5 | 91.1 | 0.0 | 91.1 | 91.1 |
| train_146 | LOO | 1 | 90.4 | 0.0 | 90.4 | 90.4 |
| | 10-fold | 10 | 87.6 | 1.7 | 84.1 | 90.5 |
| | 5-fold | 20 | 87.6 | 1.5 | 85.0 | 89.7 |
| | 3-fold | 33 | 86.5 | 2.1 | 79.5 | 89.7 |
| | 2-fold | 50 | 84.5 | 2.7 | 75.3 | 88.4 |
| train_96 | | 1 | 91.7 | 0.0 | 91.7 | 91.7 |
| test_50 | | 1 | 80.0 | 0.0 | 80.0 | 80.0 |

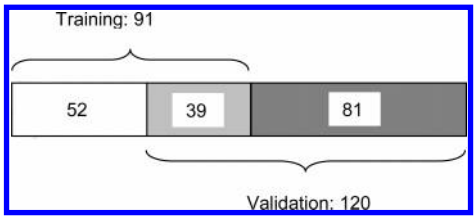



**Figure 2.** Estimation of the relative diversity of the training and validation set by counting the number of clusters for a Tanimoto coefficient of 80%.

variable selection. In the pruned tree three descriptors were selected (Figure 1). On the one hand, the acid count and basic nitrogen atom count serve as indicators for the compounds acidity or basicity. On the other hand, component 48 of the 3D autocorrelation vector giving the number of atom pairs separated by a distance interval from [5.8−5.9] Å discriminates between 2D6 and 3A4 substrates. The automatic selection by the model of size related, acidity and basicity descriptors confirms a previous finding.[50] However, the decision tree of model 2 has a potential classification issue: bulky compounds that would fit only into the CYP3A enzymatic site and that are acidic will be incorrectly predicted as a CYP2C9 substrate.

Models 3 and 4 are based on SVM models using somehow more variables and lead to a slight increase in prediction accuracy. It seems that with a predictivity of about 90% the limit of predictivity for LOO has been reached. It should be kept in mind that substrates such as those contained in the present data set may be metabolized by more than one CYP isoform.

**Selection of the Best Model.** Models 3 and 4 show the highest predictivity values. From among these two models we prefer model 3 for the following reasons: In the LOO CV process model 4 has the highest predictivity; in 2-fold CV the predictivity values are approximately equal, but the predictivity for the external data set is best with model 3. Model 3 uses the lower number of descriptors; the descriptors p$K_a$ and p$K_b$ cannot always easily or reliably be calculated. On the other hand, the descriptors used in model 3 are easily accessible by calculation methods that only need the chemical structure as input.

With the choice on model 3 we have achieved a predictivity of 89.0% in LOO CV, 84.2% in 2-fold CV, and 82.9% on an external data set of 233 compounds. High as the predictivity values of model 3 are, nevertheless model 2 also deserves attention as it only uses three descriptors and gives values of 88.4% in LOO CV, 81.6% in 2-fold CV, and 79.8% for the external data set. As this model uses a decision tree it can also easily be interpreted and utilized. In the following, we will focus on model 3 for some aspects of the discussion.

**Relevant Descriptors Confirmed/Identified in This Study.** We collected a total of 303 descriptors calculated using three different software packages. A few descriptors, related to acidity, basicity, log*P*, and size were specifically chosen based on previous works.[1] The relevance of most descriptors in our study (Table 1) for the prediction of isoform specificity was unknown. Weka provides automatic attribute selection algorithms that were applied here to find the relevant descriptors for each model descriptor set. In the following table, we report the descriptors that were selected in all four models (Table 10).

The selection was repeated 146 times using the LOO CV mode. Descriptors that were picked for all 146 CV runs are reported with a frequency of 100% (146/146). Those that were never selected are shown with a frequency of 0% (0/146). For each individual model the descriptors selected to build the models based on the whole training set train_146 are shown in boldface. For instance, model 1 is built on four 3D autocorrelation components (descriptors 176, 177, 178, and 223).

When going from model 1 to models 2, 3, and 4, we can see that only the 3D autocorrelation component 48 (descriptor # 223, 3D distance interval [5.8−5.9] Å) is still selected.

**Table 9.** Validation Results for Models 1−4

| model | aut. var. sel. | method | $n_{descr\_initial}$[a] | $n_{descr\_train}$[b] | LOO (146 cpds) | 2-fold CV (146 cpds) | validation (233 cpds) |
|---|---|---|---|---|---|---|---|
| 1 | exhaustive search | Logistic reg | 128 | 4 | 72.6 | 67.9 | 68.2 |
| 2 | exhaustive search | J48 M6 | 154 | 6 | 88.4 | 81.6 | 79.8 |
| 3 | best first | SVM E2 | 242 | 12 | 89.0 | 84.2 | 82.8 |
| 4 | best first | SVM E2 | 303 | 15 | 90.4 | 84.5 | 80.3 |

[a] $n_{descr\_initial}$ is the initial number of descriptors. [b] $n_{descr\_train}$ refers to the number of descriptors remaining after variable selection for the training set train_146.

**Table 10.** Frequency of Descriptors Picked in a LOO Cross-Validated Variable Selection

| descriptor no. (Table 1) | descriptor | occurrence [%] of descriptors picked in a LOO cross-validated variable selection | | | |
|---|---|---|---|---|---|
| | | model 1 | model 2 | model 3 | model 4 |
| 176 | 3D-$AC_{identity}$ [1.0−1.1] Å | **100** | 0 | 0 | 0 |
| 177 | 3D-$AC_{identity}$ [1.1−1.2] Å | **93** | 0 | 0 | 0 |
| 178 | 3D-$AC_{identity}$ [1.2−1.3] Å | **100** | 2 | 0 | 2 |
| 191 | 3D-$AC_{identity}$ [2.6−2.7] Å | 0 | 68 | 1 | 0 |
| 223 | 3D-$AC_{identity}$ [5.8−5.9] Å | **100** | **79** | **79** | **79** |
| 33 | $n_{acid\_groups}$ | | **100** | **100** | **100** |
| 32 | $n_{basic\_N}$ | | **100** | **100** | **100** |
| 27 | $n_{aliph\_amino}$ | | 79 | **79** | 0 |
| 29 | $n_{sec\_amino}$ | | 47 | 47 | **97** |
| 26 | $r_3$ | | 76 | **76** | **99** |
| 94 | 2D-$AC_{identity}(5)$ | | | **98** | **97** |
| 145 | 2D-$AC_{q,\pi}(3)$ | | | **100** | **100** |
| 148 | 2D-$AC_{q,\pi}(6)$ | | | **100** | **100** |
| 126 | 2D-$AC_{\chi,\pi}(5)$ | | | **97** | **95** |
| 133 | 2D-$AC_{q\sigma}(1)$ | | | **58** | **75** |
| 134 | 2D-$AC_{q\sigma}(2)$ | | | **100** | **100** |
| 116 | 2D-$AC_{\chi\sigma}(6)$ | | | **100** | **97** |
| 76 | p$K_a$ | | | | **100** |
| 77 | p$K_b$ | | | | **100** |
| 78 | log$D$ pH = 1.0 | | | | **100** |

Acidity and basicity related descriptors (descriptors 33, 32, 76, and 77) are selected with a frequency of 100%. This fact confirms previous reports about the importance of acidity and basicity in isoform specificity. On the other hand, log$P$, which is reported in previous work to differentiate CYP3A4 substrates,[1,50] is never picked in our models.

Models 3 and 4 have equally good predictive capabilities, and by large the same descriptors are selected. If the three additional descriptors needed for model 4, p$K_a$, p$K_b$ and log$D$, are dropped, their role is primarily taken up by the number of aliphatic amino groups and number of acidic groups in model 3. This simple descriptor, the number of aliphatic amino groups and the number of acidic groups, apparently suffices to reproduce the role of acid and basic groups, one more reason for selecting model 3.

Further discussion about descriptor selection is presented in the next section.

**Descriptor Role in Isoform Specificity for Model 3.** The data reported in Table 10 provide insights into the importance of individual descriptors but not about their role in model building. The decision tree presented in Figure 1 clearly shows how the three descriptors can differentiate the three isoform specificities (cf. model 2). For model 3, the best results were obtained with the help of automatic attribute selection and SVM with the exponent for the polynomial kernel set to 2.0 (SVM E2). Weka can provide SVM attribute weights that are easier to interpret when the kernel exponent is set to one (SVM E1). To investigate the role of the descriptors, we built a simple SVM E1 model using the automatically selected descriptors of model 3 (12 descriptors). SVM E1 gives similar results as SVM E2 (83.3% vs 84.2% for 2-fold CV). These descriptors and the normalized SVM weights are shown in Table 11.

**Table 11.** Selected Descriptors and Normalized SVM Weights for Three Binary Classifiers of Model 3

| no. | descriptor name | binary classifier | | |
|---|---|---|---|---|
| | | CYP2D6, CYP3A4 | CYP2C9, CYP3A4 | CYP2C9, CYP2D6 |
| 26 | $r_3$ | −2.31 | 0.00 | 0.29 |
| 27 | $n_{aliphatic\_amino}$ | 1.00 | −0.50 | −1.19 |
| 32 | $n_{basic\_N}$ | 2.00 | −0.50 | −2.19 |
| 33 | $n_{acid\_groups}$ | −0.27 | 2.00 | 0.39 |
| 94 | 2D-$AC_{identity}(5)$ | −1.24 | 0.03 | −0.08 |
| 116 | 2D-$AC_{\chi\sigma}(6)$ | −1.25 | −0.01 | −0.02 |
| 126 | 2D-$AC_{\chi,\pi}(5)$ | −2.43 | 0.01 | 0.49 |
| 133 | 2D-$AC_{q\sigma}(1)$ | 0.67 | 0.01 | −0.04 |
| 134 | 2D-$AC_{q\sigma}(2)$ | 0.83 | 0.00 | −0.14 |
| 145 | 2D-$AC_{q,\pi}(3)$ | −0.36 | 0.00 | −0.34 |
| 148 | 2D-$AC_{q,\pi}(6)$ | 0.07 | 0.00 | −0.03 |
| 223 | 3D-$AC_{identity}$ [5.8−5.9] Å | −1.50 | −0.04 | −0.09 |

To handle multiple class problems with SVM, Weka combines several two-class classifiers pairwise. Thus, the SVM attribute weights are reported for each isoform pair. For instance, for the binary classifier, CYP2C9 vs CYP3A4, the weight of 2.00 for the number of acidic functional groups (#33) is a relatively large positive number in comparison to the other weights of this classifier. A high acid count favors CYP2C9, as expected. For the two other binary classifiers the number of acidic functional groups has a much lower weight. The weight for the number of basic nitrogen atoms

**Table 12.** Compounds Missclassified Most Often

| no. | name | exp.[a] | model 1 | model 2 | model 3 | model 4 |
|---|---|---|---|---|---|---|
| 10 | cocaine | CYP3A4 | | CYP2D6 | CYP2D6 | CYP2D6 |
| 44 | quinine | CYP3A4 | CYP2D6 | CYP2D6 | CYP2D6 | CYP2D6 |
| 49 | sertraline | CYP3A4 | CYP2D6 | | CYP2D6 | CYP2D6 |
| 87 | dronabinol | CYP2C9 | CYP2D6 | CYP3A4 | CYP3A4 | CYP3A4 |
| 93 | phenytoin | CYP2C9 | | CYP3A4 | CYP3A4 | CYP3A4 |
| 102 | cyclobenzaprine | CYP3A4 | CYP2D6 | CYP2D6 | CYP2D6 | CYP2D6 |
| 103 | ebastine | CYP3A4 | CYP2D6 | CYP2D6 | CYP2D6 | |
| 106 | glibenclamide | CYP3A4 | | CYP2C9 | CYP2C9 | CYP2C9 |
| 141 | phenylbutazone | CYP2C9 | CYP3A4 | CYP3A4 | CYP3A4 | CYP3A4 |
| 142 | sufamethizole | CYP2C9 | | CYP3A4 | CYP3A4 | CYP3A4 |
| 145 | trimethoprim | CYP2C9 | CYP2D6 | CYP3A4 | CYP3A4 | CYP3A4 |

[a] Experimentally predominant isoform reported by Manga et al.[1]

(#32) is relatively high for the two classifiers CYP2D6 vs CYP3A4 (2.00) and CYP2C9 vs CYP2D6 (−2.19), again as expected. Descriptor $r_3$(#26) is a size-related descriptor. Its SVM weight is high only for the CYP2D6 vs CYP3A4 classifier. The same observation applies to component #48 of the 3D autocorrelation vector (descriptor # 223), indicating that both descriptors are used only to distinguish CYP2D6 from CYP3A4 substrates, just like in the decision tree of model 2 (Figure 1). The 2D autocorrelation descriptors #116, 126, 133, and 134 are also used to separate CYP2D6 from CYP3A4 substrates. The 2D autocorrelation descriptor #148 does not play an important role in this model as shown by its very small weights.

**Compounds Predicted Most Often Wrongly.** Tables 12 and 13 report compounds that were wrongly predicted by at least three of our models.

Cocaine (#10) is wrongly predicted as a CYP2D6 substrate by all models except model 1. Cocaine is a small polar molecule that incorporates a basic amino group that is typical of CYP2D6 substrates. This compound not only is metabolized by CYP3A4 but also is reported to inhibit CYP2D6.[51] Apparently, cocaine is too strongly bound to CYP2D6 before it can be metabolized by this enzyme. This affinity to CYP2D6 is perceived by our models.

Sertraline (#49) is wrongly predicted to be metabolized by CYP2D6 by three of our models. However, this drug is reported as a CYP2D6 and CYP2C9 inhibitor[51] as well as a substrate for the enzymes CYP2B, CYP2C19, and CYP2C9.[52] Again, the predicted affinity of sertraline to CYP2D6 is thus validated.

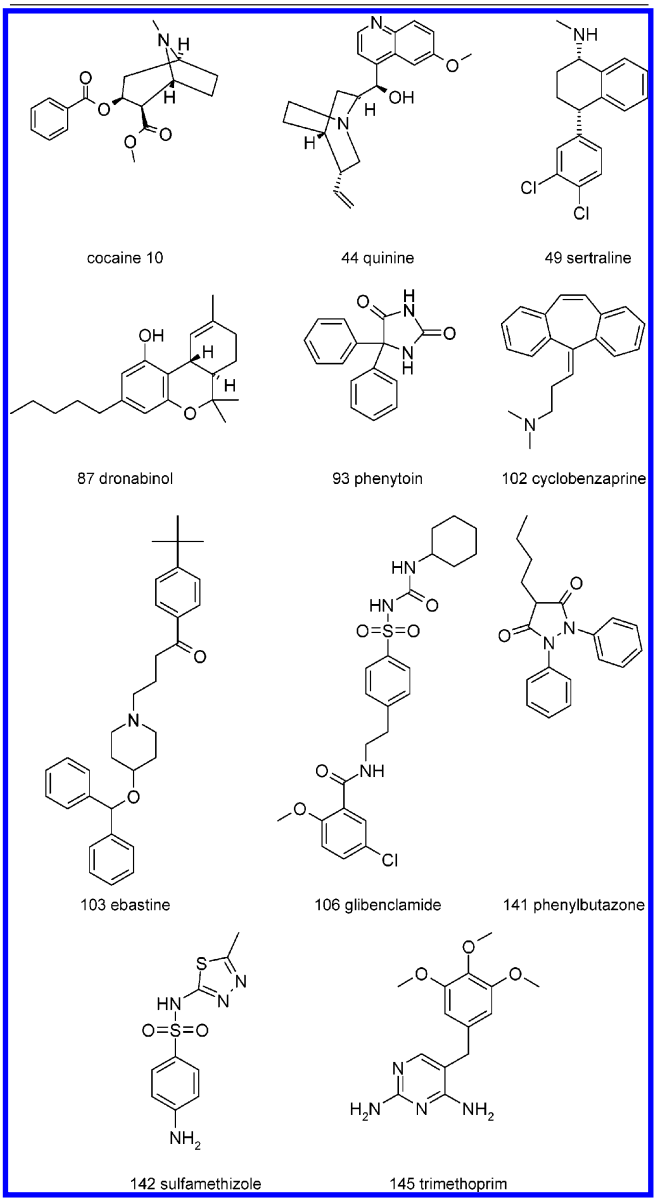
Like cocaine and sertraline, quinine (#44) is small and is a basic amine. It is predicted to be a CYP2D6 substrate by all our models. Quinine is a CYP3A4 and CYP3A5 substrate,[53] but, unlike cocaine, it is not a CYP2D6 inhibitor.[54] The predicted affinity of quinine is not correct and shows a limitation of our models.

Three of our models predicted dronabinol to be a CYP3A4 substrate probably because of the lack of an acidic functional group. Dronabinol is reported to be both a CYP3A4 and CYP2C9 substrate,[52] validating our prediction.

The drug cyclobenzaprine (#102) is incorrectly predicted by all models. This compound is reported as a CYP1A2 substrate.[51,52]

Three models predicted glibenclamide (#106) to be metabolized by CYP2C9 because of the presence of an acidic functional group. This drug is reported to be metabolized by CYP3A4[52] but also by CYP2C9,[51] again validating our prediction.

**Table 13.** Depiction of Compounds from Table 12




The two CYP2C9 substrates, phenylbutazone (#141) and sulfamethizole (#142), are incorrectly predicted as a CYP3A4 substrate by models 2, 3, and 4. Both phenylbutazone and sulfamethizole are reported to be acidic ($pK_a = 4.5$[52] and 5.4[55]). The acid count descriptor $n_{acidic\_groups}$ (Table 1), which is based on substructure count, cannot detect the acidity in

**Table 14.** Confusion Matrix Obtained for the LOO CV of Model 3

| | pred. | | | | |
|---|---|---|---|---|---|
| exp. | CYP3A4 | CYP2D6 | CYP2C9 | sum | sensitivity [%] |
| CYP3A4 | 71 | 8 | 1 | 80 | 88.7 |
| CYP2D6 | 2 | 43 | 0 | 45 | 95.5 |
| CYP2C9 | 5 | 0 | 16 | 21 | 76.2 |
| sum | 78 | 51 | 17 | 146 | |
| specificity [%] | 91.0 | 84.3 | 94.1 | | |

**Table 15.** Confusion Matrix Obtained for Prediction of the Validation Data Set with Model 3

| | pred. | | | | |
|---|---|---|---|---|---|
| exp. | CYP3A4 | CYP2D6 | CYP2C9 | sum | sensitivity [%] |
| CYP3A4 | 115 | 21 | 8 | 144 | 79.9 |
| CYP2D6 | 6 | 62 | 1 | 69 | 89.8 |
| CYP2C9 | 4 | 0 | 16 | 20 | 80.0 |
| sum | 125 | 83 | 25 | 233 | |
| specificity [%] | 92.0 | 74.7 | 64.0 | | |

phenylbutazone and sulfamethizole, hence the incorrect predictions for models 2−4.

Two CYP3A4 substrates are CYP2D6 inhibitors but not substrates as predicted (cocaine, sertraline). In further studies the models should be extended in order to discriminate between substrates and inhibitors.

Closer inspection of the literature shows that quite a few of the apparent misclassifications are actually correctly predicted either being substrates or inhibitors of the respective enzymes. Here it becomes apparent that a strict classification of quite a few drugs into being metabolized by a single CYP isoform cannot be maintained. Indeed, there is often an isoform that is the major isoform for metabolism, but other isoforms may also act as enzymes for metabolism albeit to a lesser extent.

**Confusion Matrix for the LOO of Model 3.** Misclassifications in the LOO CV affect nine 3A4 substrates (11%), two 2D6 substrates (4%), and five 2C9 substrates (24%) (Table 14).

The sensitivity is the percentage of compounds of a class which is predicted to be compounds of this class, e.g., 71 out of 80 CYP3A4 substrates are predicted to be CYP3A4 substrates, i.e., the sensitivity of the model for the class of CYP3A4 substrates is 71/80 * 100% = 88.7%. The specificity is the percentage of true positives in a set of compounds predicted to be member of this class, e.g., 16 out of 17 compounds predicted to by a CYP2C9 substrate are really substrates of CYP2C9, i.e., the specificity of the model for the class of CYP2C9 substrates is 16/17 * 100% = 94.1%.

The class of 2C9 substrates covers only 14% of the data set. Therefore, it is more challenging to predict. This is reflected by the fact that the class of 2C9 substrates shows the lowest sensitivity. Nevertheless, model 3 has the highest specificity for 2C9 substrates. All misclassified 2C9 substrates are predicted to be 3A4 substrates. Remarkably, there are no misclassifications between the classes 2D6 and 2C9.

**Confusion Matrix for the External Validation Data Set of Model 3.** The prediction of the external validation data set with model 3 gives false predictions for 40 compounds (17.2%). Among them are 29 3A4 substrates (20.1%), 7 2D6 substrates (10.2%), and 4 2C9 substrates (20%) (Table 15). The majority of conflicts are due to misclassifications between 3A4 and 2D6 substrates. One 2D6 substrate is misclassified as 2C9 substrate. All misclassified 2C9 substrates are predicted to be 3A4 substrates.

**Table 16.** Modeling Method Comparison Using the Descriptor Set of Model 3

| method | auto. var. selection | LOO | 2-fold CV |
|---|---|---|---|
| 1NN | | 80.1 ± 0.0 | 78.0 ± 3.0 |
| 1NN | BestFirst | 80.1 ± 0.0 | 79.0 ± 3.0 |
| 3NN | | 84.2 ± 0.0 | 80.7 ± 2.9 |
| 3NN | BestFirst | 86.3 ± 0.0 | 82.1 ± 2.9 |
| J48 | | 81.5 ± 0.0 | 76.8 ± 3.9 |
| J48 | BestFirst | 82.9 ± 0.0 | 78.2 ± 3.7 |
| J48 M6[a] | | 80.8 ± 0.0 | 79.3 ± 3.8 |
| J48 M6[a] | BestFirst | 87.0 ± 0.0 | 81.0 ± 3.9 |
| logistic regression | | 75.3 ± 0.0 | 69.4 ± 3.8 |
| logistic regression | BestFirst | 83.6 ± 0.0 | 75.3 ± 3.5 |
| multilayer perceptron | | 82.2 ± 2.32 | 80.0 ± 2.6 |
| multilayer perceptron | BestFirst | 84.4 ± 1.6 | 81.3 ± 2.5 |
| RBF | | 74.0 ± 1.1 | 72.7 ± 2.9 |
| RBF | BestFirst | 82.6 ± 1.6 | 79.0 ± 2.7 |
| SimpleLogistics | | 85.2 ± 1.9 | 80.5 ± 2.5 |
| SimpleLogistics | BestFirst | 84.7 ± 2.0 | 81.3 ± 2.5 |
| SVM | | 89.0 ± 0.0 | 83.3 ± 2.6 |
| SVM | BestFirst | 84.1 ± 0.3 | 83.3 ± 3.2 |
| SVM E2[b] | | 84.3 ± 0.0 | 80.3 ± 3.0 |
| SVM E2[b] | BestFirst | 89.0 ± 0.0 | 84.2 ± 3.2 |

[a] Minimum number of instances per leaf set to 6. [b] Exponent for the polynomial kernel set to 2.

For the validation data set the highest sensitivity is observed for the 2D6 substrates. On average the sensitivity for the three classes is about 3.6% lower for the validation data set than in the LOO CV. The specificity for 3A4 substrates also remains unaffected. For 2D6 substrates the specificity decreases by 9.6%, for 2C9 substrates by 30.1%. This decrease in specificity can be attributed to the fact that only 8.6% of the compounds in the validation data set are 2C9 substrates. Nevertheless, good enrichment factors are achieved for both 2D6 and 2C9 substrates. The ratio of the specificity of a class to its percentage of the entire data set gives the enrichment factor, e.g., the specificity for 2C9 substrates is 64.0% and 20 out of 233 compounds corresponding to 8.6% are 2C9 substrates. The enrichment factor for the class of 2C9 substrates in the list of compounds predicted to be 2C9 substrates is 64.0%/8.6% = 7.4. An enrichment factor of 2.5 was obtained for 2D6 substrates. The sensitivity, specificity, and enrichment factors are satisfying for this model and underline its robustness.

**Comparison of Modeling Methods.** Using the descriptor set of all 242 variables from model 3, support vectors machines gave the best CV results (Table 16). Comparable CV results are obtained for the following classification methods: *K*-nearest neighbors (1NN or 3NN), C4.5/J48, multinomal logistic regression, multilayer perceptron, RBF neural network, simple logistic regression, and SVMs. Among the classification methods investigated for this descriptor set logistic regression and RBF neural network without variable selection gave inferior results for predictions.

We found that the automatic variable selection improves the CV results for most methods that we tested for this descriptor set.

Results for methods such as multilayer perceptron, RBF, and Simple Logistics gave different results when changing the order of the compounds in the training set. Decision tree (J48) and Simple Logistics perform their own variables

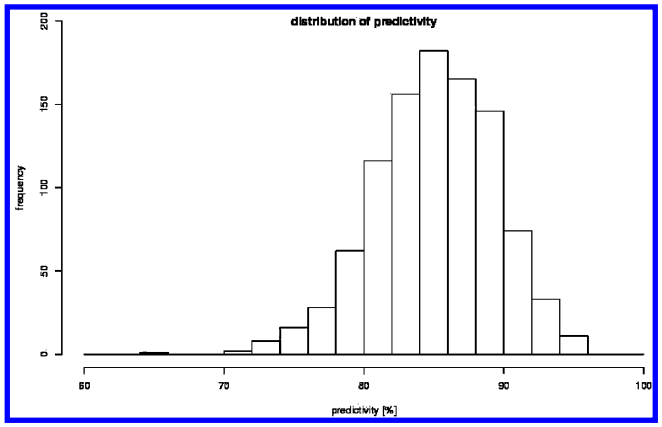



**Figure 3.** Distribution of the predictivity for 1000 randomly selected test data sets using model 3.

**Table 17.** Comparison of Prediction Using 96 Compounds as a Training Set with Descriptors of Model 3 and Variations of SVM for Model Building

| automatic variable selection | exponent | train_96 | test_50 | validate_233 |
|---|---|---|---|---|
| | 1 | 94.8 | 78.0 | 82.4 |
| | 2 | 100.0 | 70.0 | 78.5 |
| BestFirst | 2 | 90.6 | 82.0 | 82.4 |
| BestFirst | 1 | 85.4 | 72.0 | 81.1 |

selection but still gave better results in conjunction with Weka's variable selection.

As can be seen from Table 16, many different modeling methods lead to a similar predictivity, i.e., in most cases the choice of the modeling method does not have a strong influence on the model quality, whereas the choice of variables is more important than the modeling method.

**Influence of the Test Set Selection on Predictivity.** We found that, for this data set of 146 compounds, the prediction accuracy of a model heavily depends on the partitioning of the compounds into training and test set. To illustrate this point, we selected 1000 different test sets randomly (34% or approximately 50 of the 146 compounds) and computed the percentages of correct prediction using model 3. The distribution of the prediction results is shown as a histogram in Figure 3. The value ranges from 64 to 96%, with most values between 78 and 92%.

In the original study by Manga et al., the data set was divided into a training data set with 96 compounds and a test data set with 53 according to Table 1 in their paper. For the test data set an accuracy of 68.0% was achieved.[1] Using the same train/test partition, the percentage of correct predictions for testing varies from 56% (model 1) to 80% (model 4) as can be seen from Tables 3−6. We found that this percentage can also vary widely for small variations of a given model for a given set of descriptors. For instance, for model 3, we tried four variations of SVM using automatic variable selection or not and using two values for the exponent for the polynomial kernel (Table 17).

The correct predictions for test_50 set have a range from 70 to 82%. On the other hand, the variation of the correct prediction percentage for the external validation set remains small (78.5−82.4%).

## 6. SUMMARY

As can be seen from our study the choice of descriptors plays a pivotal role in modeling the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. A systematic build-up process was followed in order to come up with an optimized descriptor set. The build-up process started with an easy to calculate descriptor, the 3D autocorrelation vector reflecting the distribution of interatomic distances within a molecule. In the following, additional descriptors which carry more information related to the classification problem were added, e.g., shape descriptors and functional group counts. Automatic variable selection algorithms were used in order to decrease the number of descriptors. It was observed that variable selection slightly improves the quality of the models. The descriptors selected by variable selection confirm previous knowledge about the classification of cytochrome P450 isoform 3A4, 2D6, and 2C9 substrates.

In contrast to the choice of a suitable descriptor set, the modeling method has much less influence. Application of a comprehensive scheme of CV experiments and prediction of an external validation data set proved the robustness and reliability of the models. We achieved a LOO cross-validated predictivity of 89% for the training data set and 83% correct predictions for the external validation with our best model (model 3). It is noteworthy that the validation set is larger than the training set. Moreover, according to the cluster analysis it is also more diverse than the training set. We suspect we have just about reached the percentage of correct predictions possible as some of the compounds are metabolized by several isoforms.

Based on model 3 we showed the strong influence on predictivity of splitting a data set into a training and a test data set. Both a human and a random selection can be misleading. Algorithms to select a test data set may have the issue of being biased. A biased selection is likely to overestimate the qualtity of the model. It is our conviction that the quality of a model cannot be assessed by a single, randomly chosen test set. In fact, the distribution of the predictivity for a statistically significant number of randomly chosen test data sets should be investigated. The proportion of the different classes should be similar in both the training and the test data set. Otherwise, the predictivity of models for the training and test data set is not comparable.

## 7. CONCLUSION

The problem of predicting the isoform specificity for cytochrome P450 3A4, 2D6, and 2C9 substrates was investigated in this paper. The best model (model 3) is based on nine physicochemical descriptors which can be easily calculated with ADRIANA.Code and three functional group counts. Thorough CV of the SVM models showed 84−89% of correct predictions. The prediction of the more diverse and larger external validation data set with an accuracy of 83% underlines the validity of this model. This is a substantial improvement in comparison to the value of 68% achieved by Manga et al. for their test set. We believe we have just about reached the percentage of correct predictions possible as some of the compounds are metabolized by several isoforms.

A Web service to predict the isoform specificity is available at http://www.molecular-networks.com/online_demos.

## ACKNOWLEDGMENT

We gratefully acknowledge the BMBF for financial support in the funding initiative "Systems of Life − Systems Biology" (grant 0313080). We are grateful to Elsevier MDL Inc. for providing us with the Metabolite reaction database and MDDR database. The CACTVS toolkit was developed by Dr. Wolf-Dietrich Ihlenfeldt when he was a member of our research group. Molecular Networks GmbH assisted our work with their C++ chemoinformatics toolkit MOSES and the package ADRIANA.Code for descriptor calculation. ChemAxon made the ChemAxon Calculator available to us.

**Supporting Information Available:** Weka input files for all models plus the 3D structures for the training set (SD file format). This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Manga, N.; Duffy, J. C.; Rowe, P. H.; Cronin, M. T. Structure-based methods for the prediction of the dominant P450 enzyme in human drug biotransformation: consideration of CYP3A4, CYP2C9, CYP2D6. *SAR QSAR Environ. Res.* **2005**, *16*, 43−61.

(2) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982−992.

(3) de Graaf, C.; Vermeulen, N. P.; Feenstra, K. A. Cytochrome p450 in silico: an integrative modeling approach. *J. Med. Chem.* **2005**, *48*, 2725−2755.

(4) Crivori, P.; Poggesi, I. Computational approaches for predicting CYP-related metabolism properties in the screening of new drugs. *Eur. J. Med. Chem.* **2006**, *41*, 795−808.

(5) Fox, T.; Kriegl, J. M. Machine learning techniques for in silico modeling of drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579−1591.

(6) Schuster, D.; Steindl, T. M.; Langer, T. Predicting drug metabolism induction in silico. *Curr. Top. Med. Chem.* **2006**, *6*, 1627−1640.

(7) Bertz, R. J.; Granneman, G. R. Use of in vitro and in vivo data to estimate the likelihood of metabolic pharmacokinetic interactions. *Clin. Pharmacokinet.* **1997**, *32*, 210−258.

(8) Tredger, J. M.; Stoll, S. Cytochromes P450 - Their impact on drug treatment. *Hosp. Pharm.* **2002**, *9*, 167−173.

(9) XENIA; University of Erlangen-Nuernberg; In house CYP450 database developed at the Computer-Chemie-Centrum.

(10) MDL Drug Data Report database; MDL Inc. http://www.mdli.com/products/knowledge/drug_data_report/index.jsp (accessed Nov 2006).

(11) The PubChem project; National Library of Medicine, National Institutes of Health. http://pubchem.ncbi.nlm.nih.gov/ (accessed March 4, 2007).

(12) CrossFire Beilstein database; MDL Inc. http://www.mdl.com/products/knowledge/crossfire_beilstein/ (accessed March 4, 2007).

(13) SCIFINDER; CAS. http://www.cas.org/SCIFINDER/ (accessed March 4, 2007).

(14) Mutschler, E.; Geisslinger, G.; Kroemer, H. K.; Schäfer-Korting, M. *Mutschler Arzneimittelwirkungen. Lehrbuch der Pharmakologie und Toxikologie*; Wissenschaftliche Verlagsgesellschaft mbH: Stuttgart, 2001.

(15) Metabolite database; MDL Inc. http://www.mdl.com/products/predictive/metabolite/index.jsp (accessed March 4, 2007).

(16) ADRIANA.Code; Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com (accessed March 4, 2007).

(17) CACTVS; Xemistry GmbH. http://www.xemistry.com (accessed March 4, 2007).

(18) ChemAxon Calculator; ChemAxon, Budapest, Hungary. http://www.chemaxon.com (accessed March 4, 2007).

(19) SMARTS; Daylight Chemical Information Systems, Santa Fe, U.S.A. http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (accessed Nov 2006).

(20) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(21) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(22) Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics. Modell.* **2000**, *18*, 464−477.

(23) Gasteiger, J.; Hutchings, M. G. New empirical models of substituent polarisability and their application to stabilisation effects in positively charged species. *Tetrahedron Lett.* **1983**, *24*, 2537−2540.

(24) Gasteiger, J.; Hutchings, M. G. Quantitative Models of Gas-Phase Proton-Transfer Reactions Involving Alcohols, Ethers, and Their Thio Analogues. Correlation Analyses Based on Residual Electronegativity and Effective Polarizability. *J. Am. Chem. Soc.* **1984**, *106*, 6489−6495.

(25) Kang, K. K.; Jhon, M. S. Additivity of Atomic Polarizabilities and Dispersion Coefficients. *Theor. Chim. Acta* **1982**, *61*, 41−48.

(26) Miller, K. J. Additivity methods in molecular polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533−8542.

(27) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000; Vol. 11, pp 1−667.

(28) Hall, L. H.; Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling*; VCH: New York, 1991; Vol. 2, pp 367−422.

(29) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.

(30) 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.

(31) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. The Extent of the Relationship between the Graph-Theoretical and the Geometrical Shape Coefficients of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 714−716.

(32) Tanford, C. *Physical Chemistry of Macromolecules*; Wiley: New York, 1961; pp 1−710.

(33) Volkenstein, M. V. *Configurational Statistics of Polymeric Chains*; Wiley-Interscience: New York, 1963; pp 1−562.

(34) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(35) Csizmadia, F.; Tsantili-Kakoulidou, A.; Panderi, I.; Darvas, F. Prediction of distribution coefficient from structure. 1. Estimation method. *J. Pharm. Sci.* **1997**, *86*, 865−871.

(36) Weka: Waikato Environment for Knowledge Analysis; University of Waikato, New Zealand. http://www.cs.waikato.ac.nz/ml/weka/ (accessed March 4, 2007).

(37) Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, 2005; pp 1−524.

(38) Aha, D. W.; Kibler, D.; Albert, M. K. Instance-based learning algorithms. *Machine Learning* **1991**, *6*, 37−66.

(39) Quinlan, J. R. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*; Morgan Kaufmann: San Francisco, 2006; pp 1−316.

(40) Rumelhart, D. E.; McClelland, J. M. *Parallel Distributed Processing. Explorations of the Microstructure of Cognition*, 2nd ed.; Bradford: Cambridge, MA, London, 1986; Vol. 1, pp 318−362.

(41) Schoelkopf, B.; Sung, K.-K.; Burges, C. J. C.; Girosi, F.; Niyogi, P.; Poggio, T.; Vapnik, V. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans. Signal Process.* **1997**, *45*, 2758−2765.

(42) Landwehr, N.; Hall, M.; Frank, E. Logistic model trees. *Machine Learning* **2005**, *59*, 161−205.

(43) Platt, J. *Fast training of support vector machines using sequential minimal optimization*; MIT Press: Cambridge, MA, U.S.A., 1999; pp 185−208.

(44) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput.* **2001**, *13*, 637−649.

(45) MOSES: Molecular Encoding Structures System; Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com/software/moses/ (accessed March 4, 2007).

(46) SUBSET, a fast program for computation of a representative subset of a large data set; National Institutes of Health, U.S.A. http://cactus.nci.nih.gov/SUBSET/ (accessed March 4, 2007).

(47) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.

(48) Reynolds, C. H.; Druker, R.; Pfahler, L. B. Lead discovery using Stochastic Cluster Analysis (SCA): A new method for clustering structurally similar compounds. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 305−312.
(49) Li, W. A fast clustering algorithm for analyzing highly similar compounds of very large libraries. *J. Chem. Inf. Model.* **2006**, *46*, 1919−1923.
(50) Lewis, D. F.; Dickins, M. *Substrate SARs in human P450s;* 2002; Vol. 7, pp 918−925.
(51) Flockhart: Cytochrome P450 Drug-Interaction table. http://medicine.iupui.edu/flockhart/table.htm (accessed March 4, 2007).
(52) Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668−D672.
(53) Mirghani, R. A.; Sayi, J.; Aklillu, E.; Allqvist, A.; Jande, M.; Wennerholm, A.; Eriksen, J.; Herben, V. M. M.; Jones, B. C.; Gustafsson, L. L.; Bertilsson, L. CYP3A5 genotype has significant effect on quinine 3-hydroxylation in Tanzanians, who have lower total CYP3A activity than a Swedish population. *Pharmacogenet. Genomics* **2006**, *16*, 637−645.
(54) Donovan, J. L.; DeVane, C. L.; Boulton, D.; Dodd, S.; Markowitz, J. S. Dietary levels of quinine in tonic water do not inhibit CYP2D6 in vivo. *Food Chem. Toxicol.* **2003**, *41*, 1199−1201.
(55) *The pKa Prediction Module, Jaguar user guide*; Schrödinger Inc. http://yfaat.ch.huji.ac.il/jaguar-help/mank.html (accessed March 4, 2007).

CI700010T