articles

molecular pharmaceutics

Classification of Cytochrome P_{450} Activities Using Machine Learning Methods

Felix Hammann,[†] Heike Gutmann,[†] Ulli Baumann,[†] Christoph Helma,[‡] and Juergen Drewe*,[†]

Department of Gastroenterology & Hepatology, University Hospital Basel, University of Basel, Basel, Switzerland, and Freiburg Center for Data Analysis and Modelling, Albert-Ludwigs-University, Freiburg, Germany

Received August 30, 2009; Revised Manuscript Received October 3, 2009; Accepted October 8, 2009

Abstract: The cytochrome P_{450} (CYP) system plays an integral part in the metabolism of drugs and other xenobiotics. Knowledge of the structural features required for interaction with any of the different isoforms of the CYP system is therefore immensely valuable in early drug discovery. In this paper, we focus on three major isoforms (CYP 1A2, CYP 2D6, and CYP 3A4) and present a data set of 335 structurally diverse drug compounds classified for their interaction (as substrate, inhibitor, or any interaction) with these isoforms. We also present machine learning models using a variety of commonly used methods (*k*-nearest neighbors, decision tree induction using the CHAID and CRT algorithms, random forests, artificial neural networks, and support vector machines using the radial basis function (RBF) and homogeneous polynomials as kernel functions). We discuss the physicochemical features relevant for each end point and compare it to similar studies. Many of these models perform exceptionally well, even with 10-fold cross-validation, yielding corrected classification rates of 81.7 to 91.9% for CYP 1A2, 89.2 to 92.9% for CYP 2D6, and 87.4 to 89.9% for CYP3A4. Our models help in understanding the structural requirements for CYP interactions and can serve as sensitive tools in virtual screenings and lead optimization for toxicological profiles in drug discovery.

Keywords: QSAR; cytochrome P_{450}; machine learning; drug safety; drug design; support vector machine; artificial neural network; decision trees; *k* nearest neighbors; random forest

Introduction

The ubiquitous cytochrome P_{450} (CYP) system of hemoproteins is involved in the metabolism of a wide variety of drugs and xenobiotics. Its most common function is the oxidative and reductive degradation during phase I of drug metabolism, often as part of electron-transfer chains. It also takes part in the synthesis of endogenous compounds such as steroids and prostaglandins.[1] CYPs are a superfamily of distinct but related enzymes, with a large overlap of substrates but also considerable differences. Genetic polymorphisms, different states of induction or inhibition owing to prior exposure to certain compounds or coadministration of other drugs make the CYP system an important complicating factor in drug therapy and safety.[2]

During early drug discovery, great efforts are made to eliminate lead compounds with an undesirable toxicological profile. Special attention needs to be given to the CYP system, which is a major source of drug−drug interactions and capable of generating metabolites with potentially noxious activities. The withdrawal of the antihypertensive

* Corresponding author: Prof. Dr. Juergen Drewe, Department of Gastroenterology & Hepatology, University Hospital of Basel, Petersgraben 4, CH-4031 Basel, Switzerland. E-mail: juergen.drewe@unibas.ch. Phone: +41-61-265 3848. Fax: +41-61-265 8581.

[†] University of Basel.

[‡] Albert-Ludwigs-University.

(1) McLean, K. J.; et al. Biodiversity of cytochrome P450 redox systems. *Biochem. Soc. Trans.* **2005**, *33* (Part 4), 796–801.

(2) Vedani, A.; Dobler, M.; Lill, M. A. The challenge of predicting drug toxicity in silico. *Basic Clin. Pharmacol. Toxicol.* **2006**, *99* (3), 195–208.

10.1021/mp900217x CCC: $40.75
Published on Web 10/08/2009

Table 1. Number of Compounds (Percent of Total (n = 353)) with Given Activity for Each of the Cytochrome P_{450} (CYP) Isoforms Studied[a]

CYP	substrates	inhibitors	inducers	total active
1A2	66	33	7	88
	(18.7%)	(9.3%)	(2.0%)	(24.9%)
2D6	99	77	0	130
	(28.0%)	(21.8%)	(0%)	(36.8%)
3A4	242	78	27	264
	(68.6%)	(22.1%)	(7.6%)	(74.8%)

[a] The final column contains the number of compounds with any of these activities.

agent mibefradril from the market in 1998 after its potential for severe interactions was discovered in patients alludes to the lives and resources that may be saved if proper screening is performed beforehand.[3]

Quantitative Structure−Activity Relationships (QSAR). The fundamental concept in quantatitive structure activity relationship (QSAR) studies is the similar property principle, i.e., the assumption that similar structures have similar properties or activities.[4] The goal is to predict compound activities by analyzing their structures (*in silico*) instead of making explicit measurements *in vitro* or *in vivo*. Correlating pharmacodynamical or toxicological effects solely with structure while disregarding the molecular mechanisms at the target level may seem like an oversimplification at first, but modeling the molecular mechanisms leading to a particular pharmacological or toxicological effect is in most cases impossible, because the processes are too complex, poorly understood, or even unknown.

The goal of QSAR studies is to extract from a compound list those molecular properties that are relevant for a certain end point, usually activity or inactivity with regard to a receptor, an enzyme, or a transporter. The resulting models can be used as they are, i.e., as screening tools to predict activity status of new (untested) compounds. They may also be analyzed further to help understand the target structure or underlying mechanisms of action.

Materials and Methods

Data Set. Our data set is based on a list of FDA approved small molecule drugs (n = 1436, date of access: first of January 2008) from the University of Alberta's DrugBank database.[5] We used the DRUGDEX system (Thomson Reuters, http://www.micromedex.com, date of last access: first of May 2008) to check for interactions (as substrate, inducer, and/or inhibitor) with CYP 1A2, 2D6, and 3A4. A summary is given in Table 1. Compounds where interaction was documented were labeled "ACTIVE" whereas all other ones were labeled "INACTIVE" when the interaction status was either unknown or known to be inactive. Compounds with at least one activity or documented absence of activity were included in the final data set.

Assessment of Chemical Diversity. The structural diversity of a set of compounds can be estimated by plotting substances as points in a high-dimensional space that is spanned by the available descriptors. Similarity of two compounds increases with their proximity, and a tightly packed cluster of compounds is more homogeneous than one that is spread farther. A commonly used similarity measure is the Tanimoto coefficient.[6] Here, the similarity of two compounds i and j with k descriptors d of value X_d is expressed as

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{k} X_{di}X_{dj}}{\sum_{d=1}^{k}(X_{di})^2 + \sum_{d=1}^{k}(X_{dj})^2 - \sum_{d=1}^{k}(X_{di}X_{di})}$$

Typically, the degree of dissimilarity dis(i,j), i.e., the complement (1-sim(i,j)), is reported. It approaches 1.0 as diversity increases. To evaluate the diversity $D(M)$ of n compounds in a set M, diversity is measured for every pair of entries so that

$$D(M) = \frac{\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\mathrm{dis}(i,j)}{n(n-1)}$$

Machine Learning (ML) Methods. *k-Nearest Neighbor (kNN) Algorithm.* In *k*NN models, new instances are assigned the most common class of known instances in their immediate neighborhood.[7] During training, instances are represented as vectors in a multidimensional space. Actual calculations are only carried out during classification of a new instance (making *k*NN a lazy learning algorithm), when it is converted into a new vector and the majority class of its k-nearest neighbors is determined and returned as a result. The definition of neighborhood varies with the data studied. It can either be of a specified size using distance measures such as Hamming or Euclidean distance or be of variable size when using a fixed value for k, where the neighborhood is expanded as needed to include k instances. We chose k = 10 for our study.

Decision Tree Induction (DTI). Decision tree induction (DTI) mimics the human learning and classification process by splitting a training data set into smaller and smaller subsets, with each level having higher purity with regard to a property of interest than the one above. How purity is measured is specific to each algorithm. As learning algo-

(3) Foti, R. S.; Wahlstrom, J. L. Prediction of CYP-mediated drug interactions in vivo using in vitro data. *IDrugs* **2008**, *11* (12), 900–5.

(4) Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*, 1st ed.; Wiley: New York, 1990.

(5) Wishart, D. S.; et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34* (Database issue), D668–72.

(6) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23−24), 1046–53.

(7) Russel, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 2002.

rithms, we used classification and regression trees (CART) and the chi-squared automatic interaction detector (CHAID). The CART algorithm, introduced by Breiman in 1984,[8] uses the Gini coefficient[9] to estimate homogeneity and optimizes errors on test data during growth. By contrast, the CHAID algorithm[10] uses the χ^2-test and also performs prepruning, i.e., it halts tree growth during training before models grow too large. General growth limits for all models are a maximum depth of ten levels with at least five instances in the parent nodes and two in the child nodes. We performed no feature selection prior to learning DTI models because this is implicitly performed by the algorithms themselves.

Random Forests (RF). Taking the tree paradigm a step further, Breiman introduced random forests (RF) in 2001.[11] It creates a preset number of unpruned decision trees with randomly selected features during learning. New instances are classified by majority vote. We used forests of ten trees in our study.

Artificial Neural Networks (ANN). The concept of artificial neural networks (ANN) was first proposed in 1943 by McCulloch and Pitts.[12] In ANN models, the input vector is represented as a set of neurons with separate weighted connections to an output neuron whose state is either active or inactive, depending on the input. The network is repeatedly presented with the training data and adjusts the weights of the individual connections based on errors made until a set number of epochs (iterations) have been performed or the error converges to below a given threshold. In this constellation, ANNs are similar to linear regression models (but with a different learning paradigm) and not capable of handling nonlinear data. We therefore used an extended version, the multilayer perceptron (MLP). Here, input passes through one or more hidden layers before reaching the output neuron. Our architecture used a single hidden layer and a fixed number of 200 epochs.

Support Vector Machines (SVM). Support vector machines (SVM) are a relatively recent development. Here, each instance is represented as a point in a multidimensional space spanned by its describing features. SVM algorithms aim to find hyperplanes which separate the classes with as broad a margin as possible.[7] Because this plane cannot be deformed, conventional SVM are only applicable to linearly separable data. To work around this restriction, the so-called "kernel trick" may be used,[13] i.e., the feature space is mapped to an even higher (sometimes infinite) dimensionality by transforming the input vector using kernel functions. Several have been proposed, and we have restricted ourselves to the homogeneous polynomial function

$$K(x_i, x_j) = (\gamma x_i x_j + \text{constant})^d$$

and the Gaussian radial basis function[7]

$$K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$$

in which $\gamma > 0$ and d is kept constant.

The performance of SVM depends on the proper choice of the learning meta-parameter C (cost, a measure of the trade-off between rigidity of margins and training errors) and the kernel parameter γ. Finding the best combination of parameters can be regarded as a search in a feature space defined by C and γ, where the predictive accuracy (or, in our case, the CCR) is to be optimized. Two approaches are common: hill-climbing algorithms that start off at a random point in feature space and progress toward a local optimum, and grid searches, which evaluate every combination of parameters within a given range and resolution. Once optima are found, the evaluated region is magnified and re-evaluated with a finer resolution. Despite the greater computational expense, we chose the latter method, as it is more likely to find global optima.

Corrected Classification Rate (CCR). Usually, the accuracy of predictions is given as the ratio of hits to total number of compounds. This measure may grossly overestimate the actual quality in skewed data sets, i.e., where the members of one class greatly outnumber those of other ones. Here, we report the predictive power of each model as

$$\text{CCR} = \frac{1}{2}\left(\frac{T_N}{N_0} + \frac{T_P}{N_1}\right),$$

where T_N and T_P represent the number of true negative and positive predictions, respectively, and N_0 and N_1 the total number of negative and positive compounds in the model.

Model Validation. To avoid overfitting (i.e., creating overly complex models with very high predictive accuracy on training data by extracting too many parameters from the known data at the expense of not being able to predict unseen compounds), we used k-fold cross-validation. Here, a data set is randomly partitioned into k subsets (here $k = 10$).[14] Of these, $k - 1$ are recombined to make up a training set which is tested against the remaining subset. This process is repeated k times until all instances have served as training and test data, thereby making sure that no classes are left out.

Descriptors. Descriptors are numerical features derived from the molecular structure. In our investigation, these

(8) Breiman, L. *Classification and Regression Trees*, 1st ed.; Chapman & Hall/CRC: Boca Raton, 1984.
(9) Gini, C. *Variabilità e mutabilità*, Memorie di metodologica statistica, 1912.
(10) Sonquist, J. A.; Morgan, J. N. *The Detection of Interaction Effects*; Survey Research Center, Institute for Social Research, University of Michigan: 1964; p 296.
(11) Breiman, L. Random Forests. *Machine Learning* **2001**, *45* (1), 5–32.
(12) McCulloch, W.; Pitts, W. A Logical Calculus of Ideas Immanent in Nervous Activity. *Bull. Math. Biophys.* **1943**, *5*, 115–33.
(13) Aizerman, M.; Braverman, E.; Rozonoer, L. Theoretical foundations of the potential function method in pattern recognition learning. *Autom. Remote Control* **1964**, *25*, 821–37.
(14) Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proc. 14th Int. Jt. Conf. Artif. Intell.* **1995**, *2* (12), 1137–43.

Table 2. Overview of Descriptors ($n = 118$) Used in This Study by Origin (ChemAxon, $n = 58$; Chemistry Development Kit, $n = 61$) and Class

Class	ChemAxon	CDK
charge analysis	hydrogen bond acceptor and donor counts and sites (acceptorCounts, acceptorSiteCount, donorCount, donorSiteCount), DREIDING energy, partitioning coefficients for pH 0−14 (logD_0, logD_1, logD_2, logD_3, logD_4, logD_5, logD_6, logD_7, logD_74, logD_8, logD_9, logD_10, logD_11, logD_12, logD_13, logD_14), molecular polarizability (molPol), topological surface are (topologicalSurfaceArea), van der Waals surface area (vdwsa)	charged partial surface descriptors (PPSA-1, PPSA-2, PPSA-3, PNSA-1, PNSA-2, PNSA-3, DPSA-1, DPSA-2, DPSA-3, FPSA-1, FNSA-1, FPSA-2, FNSA-2, FPSA-3, FNSA-3, WPSA-1, WNSA-1, WPSA-2, WNSA-2, WPSA-3, WNSA-3, RPCG, RNCG, RPCS, RNCS),[38] partitioning coefficient (XlogP)
constitutional	counts of atoms, rings, and bonds (aliphaticAtomCount, aliphaticBondCount, aromaticAtomCount, aromaticBondCount, aromaticRingCount, asymmetricAtomCount, atomCount, bondCount, carboaromaticRingCount, carboRingCount, chainAtomCount, chainBondCount, chiralCenterCount, fusedAliphaticRingCount, fusedAromaticRingCount, heteroaromaticRingCount, heteroRingCount, largestRingSize, ringAtomCount, ringBondCount, ringCount, rotatableBondCount, smallestRingSize), molecular refractivity (refractivity), molecular weight (molWeight), resonant count (resonantCount)	gravitational indices (GRAV-1, GRAV-2, GRAV-3, GRAV-4, GRAV-5, GRAV-6, GRAVH-1, GRAVH-2, GRAVH-3),[39] moment of inertia along the principal axes *X*, *Y*, and *Z*, along with ratios and radius of gyration (MOMIX, MOMIY, MOMIZ, MOMIXY, MOMIXZ, MOMIYZ, MOMIR)
topological	Balaban index (balabanIndex), weighted Burden matrix (BCUTw1l, BCUTw1h, BCUTc1l, BCUTc1h, BCUTp1l, BCUTp1h),[40] Harary index (hararyIndex), hyper Wiener index (hyperWienerIndex), Platt index (plattIndex), Randic index (randicIndex), Szeged index (szegedIndex)	Kier−Hall kappa shape indices (Kier1, Kier2, Kier3), Petitjean number (PetitjeanNumber) and Petitjean indices (topoShape, geomShape), Wiener path number and polarity (WPATH, WPOL), Zagreb index (Zagreb)

include elemental analysis (e.g., atom count), charge analysis (e.g., polarizability, ion charge, topological polar surface area), geometry (e.g., number of aromatic rings, rotatable bonds), as well as partitioning coefficients, and miscellaneous other characteristics (indicators of hydrogen bonding, DREIDING energy[15]). We also included connectivity indices (Randic,[16] Platt,[17] Wiener,[18] Kier and Hall kappa shape,[19,20] and the Szeged index[21]), all of which are information-rich measures of molecular connectivity. We calculated a total of 118 descriptors (Table 2) based on the structures deposited at the DrugBank server.

Data Preparation. Compounds for which some descriptors could not be computed were removed (i.e., no missing values were allowed). Some of the algorithms are susceptible to overestimating the effects of features which are on a higher numerical scale than others (e.g., number of hydrogen bond acceptors vs molecular mass).[7] To avoid this bias, we normalized the entire data set to a range of [0, 1]. Furthermore, feature selection was performed for all algorithms except DTI and RF (which implicitly do so themselves), since they produce better results when extraneous or irrelevant information is removed *a priori*.[22] We therefore reduced the feature space to a set of 19 features with the least correlation between them using best first forward search (a greedy hill-climbing algorithm) with backtracking.[23] Lastly, classes were recoded to +1.0 for active compounds and −1.0 for inactive compounds in SVM models and 0.0 for inactive compounds in ANN models, respectively.

Software Used. ChemAxon Marvin (Marvin 5.0.4, 2008, http://www.chemaxon.com) was used for characterizing chemical structures and substructures, and ChemAxon Calculator Plugins were used for structure property calculation. Additional descriptors were calculated with the open-source

(15) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. DREIDING: A Generic Force Field for Molecular Simulations. *J. Phys. Chem.* **1990**, *94*, 8897–909.

(16) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97* (23), 6609–15.

(17) Platt, J. R. Influence of Neighbor Bonds on Additive Bond Properties in Paraffins. *J. Chem. Phys.* **1947**, *15* (6), 419–20.

(18) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69* (1), 17–20.

(19) Kier, L. B.; Hall, L. H. General definition of valence delta-values for molecular connectivity. *J. Pharm. Sci.* **1983**, *72* (10), 1170–3.

(20) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*, 1st ed.; Academic Press: San Diego, CA, 1999.

(21) Khadikar, P. V. The Szeged Index and an Analogy with the Wiener Index. *J. Chem. Inf. Comput. Sci.* **1995**, *35* (3), 547–50.

(22) 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* (4), 982–92.

(23) Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann series in data management systems, Gray, J., Ed.; Morgan Kaufmann: San Francisco, 2005.

Table 3. Corrected Classification Rates of 10-fold Cross-Validated QSAR Models of Cytochrome 1A2, 2D6, and 3A4 Interaction[a]

	CYP 1A2			CYP 2D6			CYP 3A4		
	global	substrates	inhibitors	global	substrates	inhibitors	global	substrates	inhibitors
RF	66.7	57.3	61.9	78.1	76.3	72.3	67.5	73.0	65.8
*k*NN	69.7	56.6	64.3	79.0	77.2	76.1	62.3	72.4	64.2
ANN	67.4	63.2	57.3	79.6	70.9	75.4	61.9	67.4	62.8
CHAID	**91.5**	**90.9**	**81.7**	**89.2**	**92.2**	91.6	81.4	88.6	**87.6**
CRT	78.2	78.6	70.9	87.0	89.4	**92.9**	**87.4**	**89.8**	84.7
SVM RBF	71.2	66.0	66.6	83.1	76.5	77.4	67.2	66.0	66.4
SVM polynomial	68.8	63.3	63.0	82.6	75.8	74.4	66.0	65.1	62.9

[a] RF: random forest. *k*NN: *k*-nearest neighbors. ANN: artificial neural networks. CHAID: chi-squared interaction detector. CRT: classification and regression trees. SVM RBF: support vector machines using the radial basis function kernel. SVM poly: support vector machine using the homogeneous polynomial kernel. Best values are in bold-face.

cheminformatics package Chemical Development Kit[24] (version 1.0.4, 2008, http://sourceforge.net/projects/cdk). For the calculation of certain descriptors (e.g., the set of charged partial surface area (CPSA) descriptors) 3D structures are required. These structures were generated from SMILES representations of the molecules as given in DrugBank using the Ghemical force field (http://www.uku.fi/~thassine/projects/ghemical/). With this force field, a search for lowest energy conformers was performed using the OpenBabel toolkit (Version 2.2.1, available at http://www.openbabel.org). We used SPSS (version 15.0 for Windows) for DTI and Weka (version 3.4; Waikato Environment for Knowledge Analysis, University of Waikato, Hamilton, NZ, http://www.cs.waikato.ac.nz/~ml/Weka/)[23] for *k*NN, ANN, and RF. SVM models were calculated using LIBSVM (version 2.89; http://www.csie.ntu.edu.tw/~cjlin/libsvm). Calculation of chemical diversity and grid screening for SVM metaparameters was performed with in-house software.

Results and Discussion

Chemical Diversity. We calculated the Tanimoto coefficient for the entire data set and arrived at an overall dissimilarity value $D(M)$ of 0.70. This is a high value and confirms the great structural diversity of the compounds studied. Based on this, it is fair to assume a general applicability of our models for the domain of drug-like compounds.

Feature Selection. We performed feature reduction for ANN, *k*NN, and SVM models as described. The 19 descriptors that made the final set were asymmetricAtomCount, carboRingCount, dreidingEnergy, fusedAliphaticRingCount, largestRingSize, logD_7, logD_9, rotatableBondCount, BCUTp1h, PNSA1, RPCG, RPCS, THSA, TPSA, Kier3, MOMIZ, MOMIR, WPOL, and Zagreb. This selection seems sensible, as it covers a broad spectrum of chemical features, ranging from simple counts of substructures (e.g., carboRingCount) and measures of lipophilicity (e.g., logD descriptors at physiological pH and TPSA) to complex topological indices (e.g., Kier3 and the radius of gyration, MOMIR).

Performance for Different End Points. We did not prepare models for inducers for any of the CYP isoforms because of gross underrepresentation of active compounds (Table 1). Instead, we limited ourselves to substrate and inhibitor activities and a combined end point ("global"), for which a compound is labeled as active when it shows any sort of interaction (as substrate, inhibitor, and/or inducer). If no interaction was documented, it was labeled inactive. A summary of CCRs for every end point is given in Table 3.

Cytochrome 1A2. Substrates of CYP 1A2 have been noted to be planar, aromatic, and lipophilic compounds with neutral or basic characteristics such as benzopyrene, caffeine, and propanolol.[25] Additionally, Lewis pointed out the relevance of logD partitioning coefficients at physiological pH and hydrogen bond characteristics for substances with ionizable substructures.[26] We see this reflected in the independent variables present in the most effective model (CHAID, CCR: 90.9%): acceptorSiteCount, asymmetricAtomCount, aromaticRingCount, resonantCount, FNSA3, PetitjeanNumber, dreidingEnergy, fusedRingCount, BCUTp1l, geomShape, rotatableBondCount, logD_5, logD_8, logD_9, logD_13, PSA, aromaticAtomCount, MOMIX, MOMIZ, MOMIXZ, and MOMIR.

Similarly, the global model for any type of CYP 1A2 interaction (CHAID, CCR: 91.5%), emphasizes polar features and geometric properties: vdwsa, GRAV1, fusedAromaticRingCount, RPCS, FPSA2, logD_9, WNSA3, acceptorCount, carboaromaticRingCount, WPATH, logD_13, WPSA1, logD_7, WPSA2, balabanIndex, DPSA2, logD_0, RHSA, plattIndex, geomShape, DPSA1, BCUTw1h, MOMIZ, molPol, and donorCount. More specifically, five or

(24) Steinbeck, C. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493–500.

(25) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates part 2: properties of cytochrome P450 substrates. *Drug Discovery Today* **1997**, *2* (11), 479–86.

(26) Lewis, D. F. V. Structural characteristics of human P450s involved in drug metabolism: QSARs and lipophilicity profiles. *Toxicology* **2000**, *144* (1−3), 197–203.

more hydrogen bond acceptors and at least one hydrogen bond donor appears as a prerequisite along with the presence of fused aromatic rings.

The best model for inhibition (CHAID, CCR: 81.7%) is markedly less successful. Included were the descriptors aromaticAtomCount, resonantCount, asymmetricAtomCount, BCUTp1h, logD_74, MOMIYZ, MOMIZ, PetitjeanNumber, topoShape, WNSA1, WPSA2, FPSA1, FPSA3, balabanIndex, and WPOL. The number of active compounds is also lower (half as many as in the substrate class), which suggests that a significant number of structures is classified as false negative.

Cytochrome 2D6. We achieved best results with the DTI algorithms CHAID and CRT and saw similar performance in SVM models. More than in the other end points, CYP 2D6 substrate activity seems to rely on ionizability and lipophilicity in the best model (CHAID, CCR: 92.2%). Descriptors included were logD_2, logD_3, logD_5, logD_10, logD_14, XLogP, acceptorSiteCount, donorCount, smallestRingSize, heteroRingCount, heteroaromaticRingCount, geomShape, topoShape, PetitjeanNumber, RPCG, and FPSA3. The partitioning coefficients at extreme ends of the pH range (logD_2, logD_3, and logD_14) are those descriptors which best reflect a compound's acid−base characteristics in the set of features we used. More weight is hence placed on this quality compared to, for example, the CYP 1A2 models presented above. This is in accordance with the common understanding that CYP 2D6 metabolizes its substrates via an ion-pair interaction between basic nitrogen of the substrate and aspartic acid residues at the active site.[27]

Performance for the global end point is slightly worse, with the CHAID model performing best with a CCR of 89.2%. Relevant descriptors were acceptorSiteCount, donorSiteCount, aromaticRingCount, asymmetricAtomCount, carboaromaticRingCount, smallestRingSize, logD_1, logD_5, logD_14, PPSA3, and PNSA-1. Generally, basic compounds with aromatic rings interact with CYP2D6 while hydrogen bonding behavior is of lesser importance.

The CRT model for inhibitors of CYP 2D6 had the best performance of any model in this study, relying on the following descriptors: RHSA, WNSA-1, logD_4, logD_5, logD_11, logD_12, logD_15, FPSA3, dreidingEnergy, donorCount, DPSA1, topoShape, PNSA2, BCUTc1l, and MOMIXY. Again, acid−base characteristics and lipophilicity appear, but shape constraints are of importance as well.

Cytochrome 3A4. The base for CYP 3A4's broad substrate specificity is thought to lie in its large active site where weak hydrophobic interactions determine binding.[25,28] Yap and Chen[22] reported on good models using descriptors of shape and connectivity, electronegativity, hydrophobicity, and polarizability. These were not only present in the features selected for SVM and ANN learning but also in the more successful DTI models.

For the global end point, we saw best results with the CRT algorithm (CCR: 87.4%) using hararyIndex, XlogP, logD_6, logD_11, logD_12, balabanIndex, BCUTp1h, BCUTw1l, RNCG, RPCG, geomShape, WPOL, resonantCount, THSA, aliphaticBondCount, and MOMIX. This was also the case for substrates (CCR: 89.8%) with MOMIR, XlogP, logD_0, logD_3, logD_6, logD_74, logD_8, logD_9 GRAV4, MolWeight, WPOL, RPCS, PPSA, refractivity, dreidingEnergy, geomShape, DPSA3, and aromaticAtomCount. Although these CCRs are quite good, they trail behind the values achieved for CYP1A2 and CYP2D6, whose specificity is narrower and hence more easily modeled.

For inhibitors, the CHAID algorithm produced by far the most accurate model (CCR: 87.6%), using the descriptors molPol, GRAV1, PPSA3, WNSA3, GRAVH1, MOMIXZ, WNSA1, logD_0, logD_1, logD_12, logD_13, plattIndex, PSA, acceptorSiteCount, BCUTw1l, PNSA2, resonantCount, carboRingCount, PPSA1, and rotatableBondCount.

Comparison of ML Methods. In all end points, we achieved the best results with DTI. At first glance, this seems to contradict studies such as Vasanthanathan et al.,[29] whose modeling of CYP 1A2 inhibitor activity for approximately 7400 substances using a similar portfolio of methods showed highest accuracy with SVM. DTI, however, often outperforms numerical methods (e.g., SVM or ANN) for smaller data sets, even with prior feature selection.[7] Other groups have reported accuracies of well over 90% correctly predicted instances for SVM models,[29,30] but only when applied to training sets. The same models achieve significantly lower predictive power in 5-fold CV (around 75 to 80%).

Conclusions

We present a data set of 353 compounds and their interaction status with CYP isoforms 1A2, 2D6, and 3A4. Additionally, we give cross-validated QSAR models for these activities using a set of common ML methods. Of these methods, DTI consistently outperforms its competitors. This is most probably due to the usefulness of DTI algorithms for smaller data sets.

In our survey of over 353 FDA approved compounds in DrugBank, the most frequent interactions were with CYP 3A4. In literature, this enzyme is implicated in the metabolism of more than 50% of drug compounds,[31] so the high proportion of CYP 3A4 active compounds ($n = 264$, 74.8%) can be expected. However, these account for only about 18%

(27) Langowski, J.; Long, A. Computer systems for the prediction of xenobiotic metabolism. *Adv. Drug Delivery Rev.* **2002**, *54* (3), 407–15.

(28) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates part 1: active site characteristics. *Drug Discovery Today* **1997**, *2* (10), 406–14.

(29) Vasanthanathan, P.; et al. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37* (3), 658–64.

(30) Mao, B.; et al. QSAR modeling of in vitro inhibition of cytochrome P450 3A4. *J. Chem. Inf. Model.* **2006**, *46* (5), 2125–34.

(31) Zhou, S. F. Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr. Drug Metab.* **2008**, *9* (4), 310–22.

of the compounds in DrugBank. We therefore assume that additional screening of these could uncover many other drugs with CYP 3A4 interaction potential. Arguably, these interactions could be of lesser clinical importance, either because the offending drug is not as widely used or the interaction potential is less poignant.

Furthermore, comparing our combined activity end points with the isolated activities substrate and inhibitor for each isoform, we would have expected to see significantly better CCRs in the latter, especially when the number of active compounds approaches those of the combined end point. This was seldom the case but is easily explained by the parallel membership of many compounds in different classes. For example, the antihypertensive agent metoprolol and the first generation histamine receptor antagonist promethazine both act as CYP 2D6 substrates and inhibitors, while the calcium channel blockers verapamil and amlodipine have both activities in CYP 3A4. There is significant overlap in substances and therefore in characteristics relevant for activity. In the case of CYP 1A2, models for inhibitory compounds performed markedly worse than for the combined end point, although the decrease in active substances is the same range as for the other isoforms. Most probably this is due to a large number of false negative instances in the data set.

With CCRs of 81.7 to 92.9%, our models are well suited to guide compound selection in drug discovery and also to differentiate between interactions between the different isoforms studied rather than CYP interactions in general. This is echoed in the differences between descriptors selected by the DTI algorithms for the respective end points. Models do share certain general features of drug likeness (such as lipophilicity and ionizing behavior) but rarely make use of very general descriptors such as molecular weight or measures of connectivity (e.g., the Kier and Hall kappa shape indices). The latter is often the case in QSAR studies of mass screenings of large chemical libraries, where algorithms must establish drug likeness in addition to the actual pharmacological or toxicological end point itself.

Our analysis focuses on easily calculatable descriptors and freely available modeling tools. While predictive accuracies are very high, further improvements could be expected from more sophisticated methods. Highly resolved crystal structures have been published for many CYP P_{450} isoforms which have special relevance to drug metabolism,[32] allowing the use of target-based approaches. Other groups have used methods similar to ours. In particular, Yap et al. have presented work on CYP P_{450} interactions with statistical learning methods[22,33] as have Leong et al.[34,35] Their studies show a slightly higher predictive accuracy for SVM approaches, especially when augmented with pharmacophore information. DTI models of CYP 3A4 inhibitors[33] were of lesser accuracy than the ones we present. However, most studies make use of the C4.5 algorithm[36] which is often outperformed by algorithms like the ones we employed.[37] This illustrates two important facets of QSAR modeling. First, no single method gives the best results for every data set. Second, the success of a given method is dependent on subtleties of application (setting of learning parameters, preferences of algorithms employed, etc.).

Supporting Information Available: Table of predictions for 353 compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

MP900217X

(32) Stjernschantz, E.; Vermeulen, N. P.; Oostenbrink, C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4* (5), 513–27.

(33) Yap, C. W.; et al. Application of support vector machines to in silico prediction of cytochrome p450 enzyme substrates and inhibitors. *Curr. Top. Med. Chem.* **2006**, *6* (15), 1593–607.

(34) Leong, M. K.; et al. Development of a new predictive model for interactions with human cytochrome P450 2A6 using pharmacophore ensemble/support vector machine (PhE/SVM) approach. *Pharm. Res.* **2009**, *26* (4), 987–1000.

(35) Leong, M. K.; Chen, T. H. Prediction of cytochrome P450 2B6-substrate interactions using pharmacophore ensemble/support vector machine (PhE/SVM) approach. *Med. Chem.* **2008**, *4* (4), 396–406.

(36) Quinlan, J. R., *C4.5 Programs for Machine Learning*; Morgan Kaufmann: San Francisco, 1993.

(37) Hammann, F.; et al. Development of decision tree models for substrates, inhibitors, and inducers of p-glycoprotein. *Curr. Drug Metab.* **2009**, *10* (4), 339–46.

(38) Stanton, D.; Jurs, P. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal. Chem.* **1990**, *62* (21), 2323–9.

(39) Katritzky, A. R.; et al. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100* (24), 10400–7.

(40) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (1), 28–35.