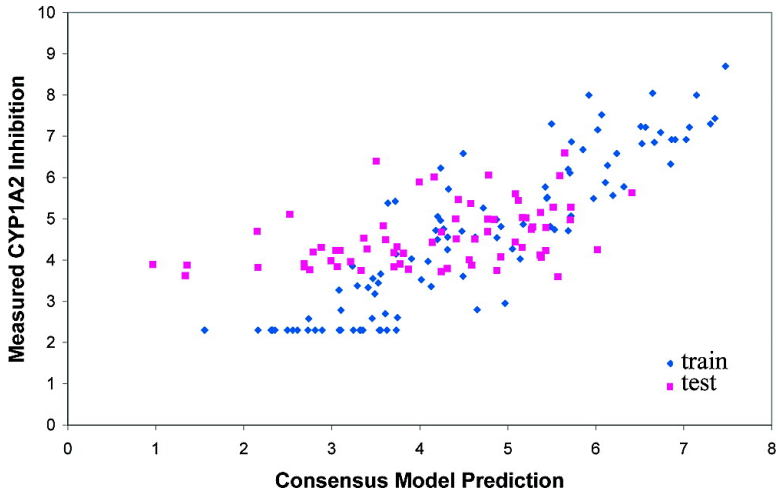
Journal of Medicinal Chemistry

Article

A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries

Kamaldeep K. Chohan, Stuart W. Paine, Jaina Mistry, Patrick Barton, and Andrew M. Davis

J. Med. Chem., **2005**, 48 (16), 5154-5161• DOI: 10.1021/jm048959a • Publication Date (Web): 12 July 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 11 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries

Kamaldeep K. Chohan,* Stuart W. Paine,* Jaina Mistry, Patrick Barton, and Andrew M. Davis

Department of Physical & Metabolic Sciences, AstraZeneca R&D Charnwood, Bakewell Road, Loughborough, Leicestershire LE11 5RH, U.K.

Received December 22, 2004

QSAR models for a diverse set of compounds for cytochrome P450 1A2 inhibition have been produced using 4 statistical approaches; partial least squares (PLS), multiple linear regression (MLR), classification and regression trees (CART), and bayesian neural networks (BNN). The models complement one another and have identified the following descriptors as important features for CYP1A2 inhibition; lipophilicity, aromaticity, charge, and the HOMO/LUMO energies. Furthermore all models are global and have been used to predict a diverse independent set of compounds. For the first time in the field of QSAR, the κ index of agreement has comprehensively been used to assess the overall accuracy of the model's predictive power. The models are statistically significant and can be used as a rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries.

1. Introduction

Cytochromes P450 are a superfamily of isoenzymes that catalyze the metabolism of a large number of compounds of both exogenous and endogenous origin. CYP1A2, which is a member of the CYP1A family of cytochrome P450s, accounts for about 12% of the total CYP content of human liver microsomes and is the major enzyme involved in the metabolism of theophylline, caffeine, imipramine, acetaminophen, and propranolol as well as the metabolism of endogenous substances such as 17β-estradiol and uroporphyrinogen III.[1,2]

Drug−drug interactions have become an important issue in health care. Many of the major pharmacokinetic interactions between drugs are due to hepatic cytochrome P450 enzymes being inhibited by concomitant administration of other drugs. The selective serotonin reuptake inhibitor fluvoxamine is a very potent inhibitor of CYP1A2. Pharmacokinetic studies have shown fluvoxamine to increase the plasma levels of caffeine when coadministered, intimating that intake of caffeine during fluvoxamine treatment may lead to caffeine intoxication.[3]

CYP1A2 is of particular importance in carcinogenesis owing to its activation of heterocyclic amines that are present in cooked meat and fish. These amines are metabolically activated by cytochrome P450 1A2 by conversion of the amino radical into a hydroxyamino group.[4] The resulting hydroxyamino derivatives are further activated by forming esters that ultimately produce DNA adducts.[5] Thus, elevated levels of CYP1A2 could enhance individual susceptibility to carcinogenesis, and modulation of this enzyme activity by CYP1A2 inhibitors could have important implications for cancer prevention. Flavonoids are a class of phytochemicals that are abundant in edible plants and tend to show CYP1A2 inhibition potential.[6] In fact, studies in vitro and in vivo have shown that some flavonoids modulate the metabolism of and disposition of xenobiotics and can contribute to cancer prevention.[7−9]

The vast majority of quantitative structure activity relationship (QSAR) studies relating to cytochrome P450 1A2 inhibitors have been based upon flavonoids and their derivatives.[10] The object of this study was to investigate the effects of an array of diverse drugs on the activity of human CYP1A2 and to elucidate structural features related to inhibitory potency. No such dataset was available to us, so we instigated an experimental campaign to generate a dataset suitable for such an investigation. Of primary importance was a large dynamic range of continuous measurements (pIC_{50}s), a diverse set of compounds to cover the drug-like space, and the properties thought to govern 1A2 inhibition.

Four statistical methods have been used in our analysis: partial least squares (PLS),[11−13] multiple linear regression (MLR), classification and regression trees (CART),[14,15] and bayesian neural networks (BNN).[16] Although these techniques employ different basic assumptions in their modeling, we would nevertheless expect to obtain complementary results in prediction. One of the advantages of regression trees and BNN is that they are able to model nonlinearity in any dataset. On the other hand MLR and PLS techniques are less abstract than regression trees and BNN, making interpretation more straightforward. Furthermore, making use of a number of techniques provides the opportunity of consensus modeling, which may have advantages over any individual model in prediction. In the fields of virtual screening, consensus modeling has been shown to produce more stable predictions than any individual model.[17,18] In the absence of measured data, model predictions can be extremely useful and provide a means of identifying molecules that may be problematic. QSAR models are more than a literature curiosity, and successful ones offer the potential to be used as a virtual screen to filter design targets before synthesis. However,

* To whom correspondence should be addressed. Phone: +44 (0)-1509 64 4882. Fax: +44 (0)1509 64 5576. E-mail: Kamaldeep.Chohan@astrazeneca.com and Stuart.Paine@astrazeneca.com.

10.1021/jm048959a CCC: $30.25
Published on Web 07/12/2005

Table 1. Training Set Statistics for Each Model

model	model r^2	rmse
PLS	0.72	1.0
MLR	0.71	1.0
CART	0.84	0.7
BNN	0.72	1.0

as has recently been shown by Stouch,[19] apparently reasonable models can fail to give useful predictions when tested on compounds structurally different from the training set. We therefore wished to investigate the performance of these models in true prediction, and also to evaluate the dependence of predictivity on the distance of the predicted compound to the model space.[20]

2. Results and Discussion

Based on the rmse of the training set, all models except for CART have the same rmse (error = 1 log unit; Table 1). The CART model has the lowest rmse = 0.7 log unit.

2.1. Partial Least Squares Analysis. A significant PLS model based on pIC_{50} data has been produced. The 2 component PLS model incorporates 17 *x*-descriptors and has an $r^2 = 0.72$ and a $q^2 = 0.67$ ($n = 109$, leave-1-out). This is a good model and shows high predictivity in an internal cross-validation study. The most important variables describing CYP1A2 inhibition were lipophilicity and aromaticity (number of aromatic carbons). The model predicts that decreasing lipophilicity and aromaticity should decrease CYP1A2 inhibition. Furthermore, increasing charge, dipole moment, M3M (i.e. moment of inertia along the third principal axis of a molecule), and the difference between the HOMO and LUMO energy decreases CYP1A2 inhibition.

2.2. Multiple Linear Regression Analysis. A five-term MLR model has been produced. The model fit details are similar to the PLS case. The MLR model has an $r^2 = 0.71$ (r^2 adjusted = 0.70). Each of the terms is statistically significant. The effect of each variable is summarized in eq 1:

$$pIC_{50} = 4.457 + (0.357 \times \text{ACDLogD7.4}) - (0.952 \times \text{DE}) - (0.0989 \times \text{dipole moment}) + (0.113 \times \text{aromaticity}) - (0.0808 \times \text{M3M}) \quad (1)$$

Consistent with the PLS model, lipophilicity (i.e. ACDLogD7.4) and aromaticity are positively correlated with CYP1A2 inhibition and dipole moment, M3M, and DE (i.e. the difference between the HOMO and LUMO energy) are negatively correlated with CYP1A2 inhibition.

2.3. Classification and Regression Trees Analysis. Regression analysis carried out within CART produces a significant model based upon the pIC_{50} data and has $r^2 = 0.84$ and rmse = 0.70. Directly understanding the properties that have the largest influence upon the inhibition of CYP1A2 from this CART regression model can be rather challenging. The way these authors have approached the problem is to investigate each of the individual 15 trees, which make up the final regression analysis. It is then possible to look at a frequency distribution of the most important properties in each of these trees and from this conclude which descriptors are having the most influence on the model overall. Figure 1a shows a frequency distribution plot of the most significant descriptors in each of the 15 trees used to build the regression model. As can be seen, aromaticity is the most significant variable in five out of the 15 trees. MMSPECV1, which is a measure of positive electrostatic potential on the van de Waals surface area, is the next most frequent variable, occurring in 3 of the 15 trees as the most important variable. ACDLogD7.4 and the HOMO energy both occur twice as the most important variable. Figure 1a suggests that the CART regression model shows concordance with the PLS, MLR, and BNN models, as all models are identifying aromaticity and lipophilicity as playing major roles in the inhibition of CYP1A2. Increasing aromaticity and or increasing lipophilicity will increase the inhibition of CYP1A2.

Figure 1b shows the frequency distribution for all descriptors that fall within the top five most important descriptors for each of the 15 trees, which make up the regression model. In this distribution the most frequent is MMSPECVD, which is a measure of the negative electrostatic potential on the van de Waals surface area. This would indicate that, while this descriptor is not frequently the single most important descriptor, it does contribute significantly to all the trees used to build the regression model. Aromaticity and ACDLogD7.4 are again highly significant descriptors in all the trees used in the CART regression analysis.

2.4. Bayesian Neural Networks Analysis. The statistics of the BNN model are similar to the PLS and MLR models: $r^2 = 0.72$ and an rmse = 1.0. The BNN model incorporates six descriptors; most of these are similar to the descriptors in the PLS, MLR, and CART models (i.e. aromaticity, lipophilicity, M3M, and the HOMO/LUMO energies). The BNN model has a further two descriptors for solvent accessible surface area and the number of acidic groups likely to be ionized at pH = 7.4.

Generally, all four models suggest that lipophilicity, aromaticity, charge, and HOMO/LUMO energies are important features describing CYP1A2 inhibition. Interestingly, these descriptors are also found to be important in describing CYP1A2 substrates.[21]

2.5. Consensus Model. The different statistical approaches for modeling CYP1A2 inhibition allow the use of consensus modeling. In this paper, the consensus prediction of a compound is simply the average prediction from all models. A consensus model may be better than any individual model simply because an extreme prediction by one model may be a subtle outlier or a reasonably predicted compound by the other models. Figure 2 shows the observed versus predicted values using the consensus model for both the training set (blue diamonds) and the independent test set (pink squares). Compounds that had their pIC_{50} pinned at 2.30 because of no apparent inhibition were included in the training set. The variance of the prediction for these pinned compounds is similar to the variance for predictions over the full dynamic range of measured pIC_{50}s. Furthermore, removal of these points from the training set does not significantly affect the outcome of the models. Therefore, inclusion of these points in the training set is reasonable and ensures that the training set covers the oral drugs' property space. The consensus model

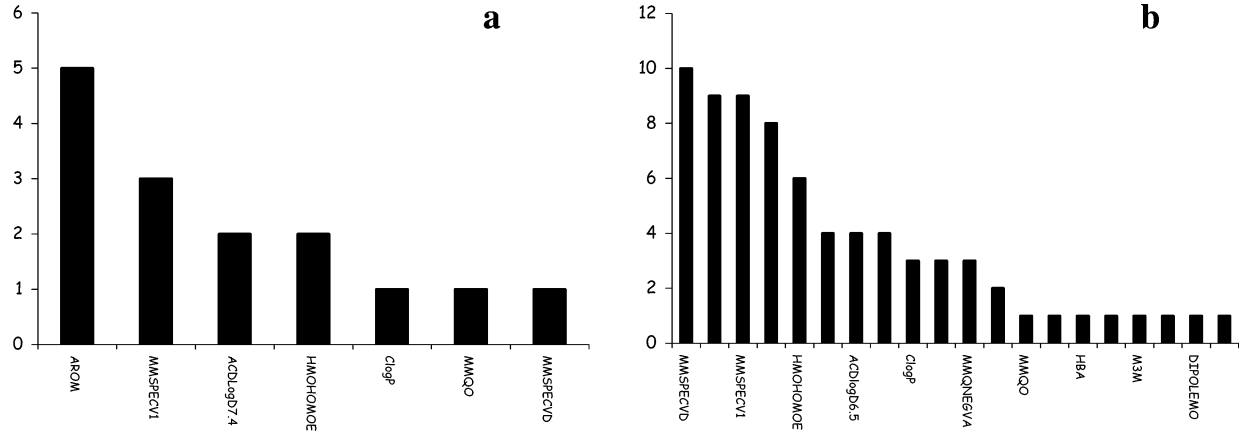


Figure 1. (a) Frequency distribution profile for descriptors that occur as the most important descriptor in each of the 15 trees used to make the regression model. (b) Frequency distribution profile for all descriptors that occur within the top five descriptors of each of the 15 trees used to make the regression model.

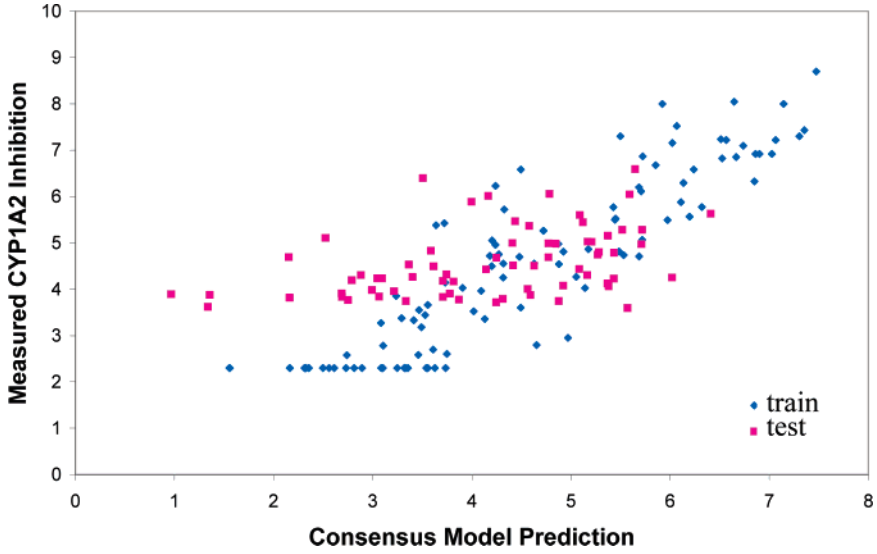


Figure 2. Measured CYP1A2 inhibition (pIC_{50}) versus the consensus model prediction.

Table 2. Statistics for the Oral Drug Test Set (68 Compounds)[a]

model	r^2	mean error	r^2_{bf}	rmse
PLS	−2.18	−0.58	0.16 (<0.001)	1.3
MLR	−2.59	−0.41	0.17 (<0.001)	1.4
CART	−0.92	−0.20	0.20 (<0.001)	1.0
BNN	−2.29	−0.43	0.14 (0.002)	1.3
consensus	−1.58	−0.41	0.19 (<0.001)	1.2

[a] F significance is in parentheses.

(rmse = 0.84) is a better model than the PLS (rmse = 1.0), MLR (rmse = 1.0), and BNN (rmse = 1.0) models but is weaker than the CART (rmse = 0.70) model. However, to assess the model predictivity it is best to evaluate the rmse of an independent test set.

2.6. Analysis of Model Predictions. If we consider the rmse in the test set for the 68 compounds that do have continuous measurements, then the CART model (rmse = 1.0) is better than the consensus model (rmse = 1.2), which is better than all other models (PLS and BNN rmse = 1.3 and MLR rmse = 1.4; Table 2 and Figure 2). However, the rmse of these models is worse than the standard deviation of the measured pIC_{50}s in the test set. Furthermore, the rmse values of the models are no better than the rmse obtained if the mean pIC_{50} value of the training set is used to predict the 68 compounds in the test set. This is reflected by the negative r^2 around the line of unity, however the r^2_{bf} (around the line of best fit) is both positive and statistically significant (Table 2). This suggests that there is a bias in our predictions, and this is also reflected in the mean error, implying that our models are underpredicting for this specific test set. Nevertheless, the models can rank compounds based on CYP1A2 inhibition, as there is a statistically significant relationship between measured and predicted. On the other hand, models could have been built and validated by partitioning the training set of 109 compounds into 87 training and 22 test set compounds (i.e. 80:20 split). However, while a model may be built on a dataset produced in a single screen from a single laboratory, or from a literature dataset, its success is often judged by its success in prediction as measured in a similar screen run in another laboratory. Hence, the authors decided to use the 249 oral drugs that had their pIC_{50}s determined at a different AstraZeneca site as their independent test set. Furthermore, this test set is a challenging test in that it only covers the middle range of the training set compounds in terms of the pIC_{50}s (Figure 2). However, in the screening of a new chemical series the dynamic range is likely to be unknown.

If compounds in the test set are similar in properties and or structurally to the training set compounds, then the model is likely to predict these compounds reasonably well. We have assessed the Euclidean distance of a compound i to a compound j belonging to the training set[20] (eq 2),

$$\mathrm{Dist}_{i,j} = \sqrt{\sum_{d=1}^{P} \frac{(D_{dj} - D_{di})^2}{\sigma_d^2}} \qquad (2)$$

where D_{dj} is the value of the dth descriptor of the jth compound of the test set and D_{di} is the value of the dth descriptor of the ith compound in the training set. For a compound x, there are as many distances as there are compounds in the training set. This set of distances forms a vector $D_{x,\mathrm{train}}$. For a compound, the distance to its nearest neighbor in the training set is defined as the minimum value of the vector $D_{x,\mathrm{train}}$.

For the distance to the center of the training set, D_{di} is replaced by the mean of descriptor D, x_d. The normalized distance ND_i to the center of the training set or to the nearest neighbor in the training set is expressed in eq 3,

$$\mathrm{ND}_i = \sqrt{\frac{15}{d}}(D_{i,\mathrm{train}}) \qquad (3)$$

where d is the number of descriptors of the model.

Similar measures have also been used previously;[22] here the authors find that prediction error is related to the similarity between a test set compound and the training set compounds. The assumption is that as compounds move further away from the property space of the model then the error in prediction will increase. To evaluate this, the 68 test set compounds which have reported pIC_{50} values were used, and in each case 2 types of distance to model calculations were performed for each compound: (i) Euclidean distance to nearest member of training set and (ii) Euclidean distance to center of training set.[20] The normalized distances ND_i were binned as >1 and <2, >2 and <3, >3 and <4, and >4 (Figure 3), and the rmse (based on the consensus prediction) was determined for each bin. We find that the test set is dissimilar from the training set as there are no compounds with a distance to model of less than 1. The error in prediction of the test set shows a relationship to both the Euclidean distance to the center of the training set and to the nearest neighbor in the training set. This analysis suggests that distance to the model space may be a critical factor in determining the accuracy of the predictions of the 68 compounds from the training set. The relationship between distance and the precision of predictions based on the test set may be used to give confidence in future predictions of compounds at similar Euclidean distance to the training set.

Although the models are based on continuous pIC_{50} data, the model predictions have also been used in a classification sense because so many compounds in the test set were reported as out of range data. Nevertheless, the classification approach has its own limitations. For instance, we have used a cutoff of $\mathrm{pIC}_{50} = 5$: although this is perfectly justified in terms of progressing good compounds, by using an absolute cutoff we have increased the likelihood of misclassifications. In the classification of the 249 oral drugs test set, all models give similar accuracy measures as calculated from their confusion matrix. The observation that these four CYP1A2 models show similar predictivity may suggest that the modeled biological property could be linear in nature. For conciseness we have presented accuracy measures calculated from the consensus model confusion matrix (Tables 3 and 4). The per class measures of sensitivity and specificity do not address the problems of the prevalence of the test set. However, the positive and negative predictive values may be more useful as they are less sensitive to prevalence. The positive and negative predictive power of the model are 23% and 96%, respectively, and statistically better than expected from chance based upon a knowledge of either the training or test set prevalences. Furthermore, the model is better than assigning the predictions to only one class (i.e. $\mathrm{pIC}_{50} < 5$). In Table 4, the probabilities in parentheses represent the probability that the model prediction is no better than the prevalences. The correct classification rate (both classes) for the model (83%) is better than chance based upon the training set prevalence (56%) but no better than chance based upon either the test set prevalence (87%) or assigning to one class (93%). However, the correct classification rate is a naive measure of overall accuracy, as it does not take into account that some or all of the apparent classification accuracy could be due to chance. The κ index, which does take chance into account, calculated for our model is

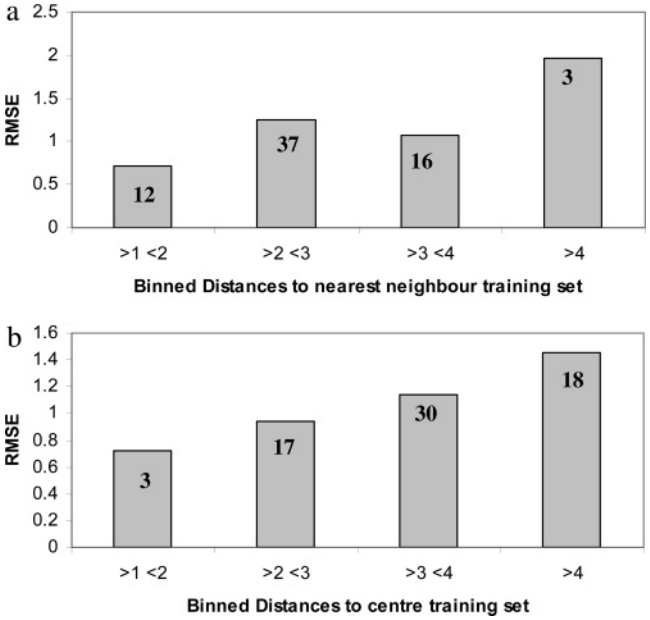


Figure 3. The rmse error against distance to model for (a) to nearest neighbor and (b) to center of training set.

Table 3. A Confusion Matrix for the Oral Drug Test Set[a]

		Measured	
		pIC50 > 5	pIC50 < 5
Predicted	pIC50 > 5	10	34
	pIC50 < 5	8	197

[a] Data generated based on the consensus model.

Table 4. Consensus Model Classification Statistics for the Oral Drug Test Set[a]

measure	model prediction	chance: based on training set prevalence (40:60)	chance: based on test set prevalence (7:93)	assign to one class only ($pIC_{50} < 5$)
correct classification rate	83%	56% (<0.001)	87% (0.96)	93% (>0.99)
sensitivity	56%	40% (0.09)	7% (<0.001)	0% (<0.001)
specificity	85%	60% (<0.001)	93% (>0.99)	100% (>0.99)
false positive rate	15%	40% (<0.001)	7% (>0.99)	0% (>0.99)
false negative rate	44%	60% (0.09)	93% (<0.001)	100% (<0.001)
positive predictive power	23%	7% (0.001)	7% (<0.001)	7% (<0.001)
negative predictive power	96%	93% (<0.001)	93% (0.04)	93% (>0.99)
κ	0.25	0.00 (0.04)	0.00 (0.01)	0.00 (0.05)

[a] All data has been generated based on the confusion matrix. Probabilities are in parentheses.

Table 5. Interpretation of the κ Statistic

κ	interpretation
<0	no agreement
0.0−0.19	poor agreement
0.20−0.39	fair agreement
0.40−0.59	moderate agreement
0.60−0.79	substantial agreement
0.80−1.00	almost perfect agreement

statistically different from zero, unlike our prediction scenarios based upon training, test set, and assign to one-class prevalences. In the chance scenarios κ is zero, implying no agreement between prediction and measurement, and for the model a κ of 0.25 means that we have a fair agreement (Tables 4 and 5). These accuracy measures are highly encouraging as the test set was independent of the training set and measured in a different AstraZeneca site. Ideally a test set with no prevalence in either class would be a more definitive test of the model; in reality, the prevalence is likely to be unknown for the screening of a new chemical series.

To date the authors are not aware of any global diverse CYP1A2 literature QSAR models. The literature does contain series-based models (e.g. flavonoids),[10] and it would have been valuable to rebuild these models and predict our oral drug test set: this would have been a real measure of how our model compares with the literature cases. Unfortunately, this was not possible because we do not have access to descriptors that make up the literature models.

3. Conclusions

We have used 4 statistical approaches in order to understand CYP1A2 inhibition potential —although these methods differ overall, nevertheless we have obtained similar results (e.g. in terms of the descriptors for each model and the performance of each model in predicting an independent test set). The PLS and MLR methods have been extremely useful in terms of understanding the influence and effect of a descriptor on CYP1A2 inhibition. These in turn can be used to guide chemistry and to move away from compounds that have CYP1A2 inhibition. These models suggest that lipophilicity, aromaticity, charge, and HOMO/LUMO energies are important features describing CYP1A2 inhibition. In all cases the models are global and may be used to predict a diverse range of compounds, which is a potential limitation of other literature CYP1A2 models. These models can be used as a rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries.

4. Materials and Methods

4.1. Chemicals. All chemicals and reagents used were of the highest available commercial grade. α-Naphthoflavone, 1-aminobenzotriazole, anthracene, beclomethasone, β-carotene, bupropion, busulfan, capsaicin, chloramphenicol, cholic acid, cinoxacin, citral, curcumin, cyclosporine, d-limonene, dantrolene, dextromethorphan, diclofenac, digitoxin, digoxin, diltiazam, dipyridamole, disopyramide, disulfiram, ellipticine, enoxacin, epigallocatechin, erythromycin, fenfluramine, flavone, flurbiprofen, furafylline, furazolidone, furosemide, glucosamine, hydroxyurea, ibuprofen, kanamycin, ketoconazole, labetalol, lidocaine, mepenzolate bromide, methimazole, methocarbamol, methylene blue, mexiletine, morphine sulfate, nadolol, niacin, nifedipine, norfloxacin, fluvoxamine, phenethylisothiocyanate, propafenone, propofol, propranolol, pyrene, pyridoxine, quinacrine, resveratrol, tamoxifen citrate, tannic acid, tetracycline, tolbutamide, tryptamine, valproic acid, verapamil, ethoxyresorufin, and β-nicotinamide adenine dinucleotide phosphate, reduced form (β-NADPH), were purchased from Sigma Chemical Co. (Poole, U.K.). Galangin was purchased from Aldrich Chemical Co. (Gillingham, U.K.). Omeprazole, serotonin, cimetidine, and the in-house compounds 1−15 (purity > 99%) were synthesized at AstraZeneca Charnwood (Loughborough, U.K.). *Escherichia coli* coexpressing P450 1A2 and human NADPH, P450 reductase were purchased from Cypex (Dundee, U.K.). Previous studies have demonstrated that supplementation with cytochrome b_5 is not required for this system: kinetic parameters similar to both other recombinant systems and human liver microsomes for the CYP1A2 isoform have been reported.[23,24]

4.2. Automated Ethoxyresorufin O-Deethylation Inhibition Assay. Ethoxyresorufin O-deethylation was used as the probe reaction for CYP1A2 and was based on an automated assay previously described.[24] Fourteen compounds, including fluvoxamine (positive control) at six concentrations, were screened per 96-well plate per run. Test compounds (e.g. 5 mM) in DMSO were diluted in water by a robotic sample processor (RSP), giving a range of concentrations (e.g. 250 to 1 μM) with the DMSO constant at 5% (v/v). Stocks were diluted 1:10 into the incubation to give an appropriate range of concentrations for each test compound. Each incubation contained 60 μL of NADPH (1.6 mM) and 100 μL of protein (0.1−0.5 mg/mL final concentration) to give 15 pmol of enzyme/mL, and 20 μL of test compound in 5% DMSO and 20 μL of ethoxyresorufin (6 μM) in 2% DMSO were added to start the reaction. Thus the final concentration of DMSO in the incubation was 0.7%. An incubation containing DMSO alone allowed calculation of control activity. Production of resorufin (λ_{ex} 544 nm, λ_{em} 590 nm) was measured over 15 min (33 readings) on a fluorescence plate reader (f_{max}; Molecular Devices Co. Sunnyvale, CA). All measured data represent means from at least triplicate determinations.

4.3. Data Analysis. Kinetic parameters were determined by linear or nonlinear regression using Microsoft Excel (Redmond, WA), Microcal Origin 6.0 (OriginLab Corporation, Northampton, MA), or WinNonLin 3.1 (Pharsight, Mountain View, CA). IC_{50} values were determined by linear transformation within Microsoft Excel.

4.4. Datasets. The training set consisted of 109 compounds; 87 of these were commercially available oral drugs, and 22 were in-house compounds (see Supporting Information).[25−34] A mean pIC_{50} value was measured for all of the in-house compounds and 59 of the commercially available ones. The average standard deviation of the measured pIC_{50} data for triplicate determinations is 0.2. The pIC_{50} values for the remaining 28 oral drugs were obtained from the literature.[25−34] Many of these pIC_{50} values were estimated from the literature K_i value based upon the assumption that inhibition was competitive (i.e. $IC_{50} = 2K_i$). If inhibition is noncompetitive, then the IC_{50} will equal K_i and therefore a 2-fold error would arise as a result of assuming competitive inhibition. Consequently, the QSAR models that have been built represent a combination of literature as well as in-house data. For compounds that showed no sign of inhibition at the top concentration of 2 mM, a value of 5 mM was set as the IC_{50} ($pIC_{50} = 2.3$) and used for the purposes of model building. While this is unusual in developing a quantitative model, it ensures that the training set contains compounds that represent the drug-like space of marketed oral drugs. One hundred twenty-three descriptors, which broadly describe topological, geometrical, and electronic features of molecules, were calculated using an in-house descriptor generator engine. These descriptors have been described in detail elsewhere.[20,35,36] In this context the authors define global as a dataset comprising compounds of a wide structural diversity and a good inhibitory range; correspondingly, the models should be able to predict across structural classes. The diversity in property space was ensured using hierarchical clustering with a database of 594 marketed (in the USA) oral drugs created from an analysis of the USA pharmacopoeia *Physicians' Desk Reference 1999*.[37,38] In hierarchical cluster analysis the Euclidean distance (i.e. this makes use of a straight-line distance as a measure of dissimilarity) was used to calculate clusters from 6 principal components that contain approximately 80% of the information from the original 123 x-variables; the 6 principal components were calculated via SIMCA-P (version 8; Umetrics).[39,40]

4.5. QSAR Modeling. The pIC_{50} inhibition data was modeled using 4 statistical packages: PLS, MLR, CART, and BNN. For each modeling method the same descriptor set of 123 descriptors was used. Where possible, automated variable selection was used to extract variables that are likely to be key for inhibiting CYP1A2. To test each model a large independent test set was used. This test set contains 249 oral drugs that were measured at a different AstraZeneca site than the training set measurements. This cross-site validation shows how a QSAR model might be used in a large pharmaceutical company. While a model may be built on a large dataset produced in a single screen from a single laboratory, or from a literature dataset, its success is often judged by its success in prediction as measured in a similar screen run in another laboratory.

4.5.1. Partial Least Squares. To obtain an optimal PLS model for CYP1A2 inhibition, the program GOLPE[41] was employed with PLS modeling to select key variables describing CYP1A2 inhibition (pIC_{50} data). The x-matrix prior to automated variable selection consisted of 123 descriptors. D-optimal preselection was then used to remove 20% of variables that have little significance in the PLS model (i.e. variables which contain little or redundant information): this reduced the x-matrix to 98 variables. Full factorial selection was then used to filter down to key variables describing CYP1A2 inhibition: this filtering process involves assessing the contribution of each variable to the predictivity of the PLS model. SIMCA-P (Umetrics) was then used to rebuild the PLS model using variables selected by GOLPE: 17 variables were selected in total. Cross-validation procedures gave rise to q^2 values, and the appropriate number of principal components was determined when the q^2 reached a maximum. Randomization tests are a useful indication of model robustness. To test the robustness of the PLS model a randomization test was performed 999 times (within SIMCA-P) on the initial observed y-data. No randomized case was statistically better than the initial model r^2 or q^2, implying that our model is better than random.

4.5.2. Multiple Linear Regression. The initial 123 descriptor matrix is likely to contain similar descriptors, which in MLR could cause cross-correlation problems (i.e. descriptors encoding a similar molecular property). To elucidate this initial problem and therefore make the MLR modeling more tractable we used the 17 descriptors from GOLPE variable selection as our initial starting point for MLR modeling. An MLR model was obtained within JMP 4 (version 4.0.2).[42] A procedure of alternating the forward addition and backward elimination steps reduced the 17 GOLPE descriptors to 5.

4.5.3. Classification and Regression Trees. The program CART was used to perform the regression tree analysis. The CART methodology employs binary recursive partitioning, and in this model a consensus of 15 regression trees was found to be optimal when combined using bootstrapping aggregation (Bagging). The Gini algorithm together with least absolute deviation regression was used throughout this work. No misclassification costs were used in this analysis, and priors were set as equal.

4.5.4. Bayesian Neural Networks (BNN) and Automatic Relevance Determination (ARD). BNN models are less susceptible to overtraining and overfitting compared to classical neural networks.[43] Several publications have used BNN techniques for building QSAR models and have been especially successful when applied to ADMET modeling.[44−47] For example, Sorich et al. have shown the BNN approach to produce superior models compared with linear techniques when applied to the mapping of phase II metabolism.[47]

A BNN model was produced using scripts in Perl language written by P. Bruneau,[20] coupled with an automated routine for variable selection written by R. Neal.[48] Prior to feeding the data into the BNN both the descriptor vectors and the dependent variable were scaled to give a mean equal to 0 and a standard deviation equal to 1. The protocol followed to give a BNN model has been described by Bruneau.[20] This paper should be consulted for a full discussion on training parameters. For 105 cases, the architecture of the initial BNN consisted of 122 nodes in the input layer, 2 nodes in the hidden layer, and 1 node in the output layer. Each node in the input layer was connected to all the nodes in the hidden layer and the output layer node. In addition a bias node linked to all the hidden nodes is added to the input layer and a bias node linked to the output node is added to the hidden layer. The transfer function is a hyperbolic tangent. The BNN was used together with automatic relevance determination (ARD)[49] to select the most relevant descriptors to generate the most parsimonious model. ARD starts by initially building the most complete BNN model utilizing all the descriptors. The BNN model was trained for 500 cycles. The ARD then adds to each input unit a hyperparameter, which controls the magnitudes of the weights of the connections of that input unit. As training proceeds, the weights associated with irrelevant descriptors are forced to small values, while the weights of important variables are allowed to take high values. ARD parameter takes a value, which hardens the process as it becomes smaller. After the network is trained, the distribution of weights associated with every descriptor is analyzed, and only the descriptors with reasonably large weights are retained for the next step. Descriptors with an associated sum of weights less than 1% of the maximum sum of weights are discarded. The process of training (500 cycles) and removal of descriptors is looped until no further descriptors are removed in 5 successive iterations; the ARD is then incremented to its next value. The ARD parameters available were 5, 2, 1, 0.5, 0.2, 0.1, and 0.05. The process continues until it is impossible to remove any more descriptors without degrading the performance of the resulting model. Table 6 summarizes the training and ARD parameters together with the BNN r^2, where ρ is the ratio of the number of cases to the total number of connections in the network. The final BNN model was obtained by training for 1000 cycles, and the number of input nodes was equal to 6 as selected by

Table 6. Summary of BNN and ARD Parameters

step	no. of cases	no. of input nodes	no. of hidden nodes	ρ	ARD param	r^2	descriptors left for next step
1	105	122	2	0.4	5	0.79	21
2	105	21	2	2.2	5	0.75	21
3	105	21	2	2.2	5	0.76	20
4	105	20	2	2.3	5	0.76	20
5	105	20	2	2.3	5	0.79	20
6	105	20	2	2.3	5	0.73	20
7	105	20	2	2.3	5	0.79	20
8	105	20	2	2.3	5	0.79	20
9	105	20	2	2.3	5	0.79	20
10	105	20	2	2.3	2	0.73	6
11	105	6	2	6.2	no	0.72	6

Actual

Predicted	+	-
+	a	b
-	c	d

Figure 4. A confusion matrix. In this table, a, b, c, and d represent true positive, false positive, false negative, and true negative, respectively.

ARD. The final model contained 6 relevant descriptors, and the last 200 cycles were used to give the final predictions in the unscaled form. The last 200 cycles were used because this gives a representative sampling of the BNN training process.[48] For most literature models, the stopping criteria for a BNN involves training the net until the log of the evidence is a maximum. For our implementation, the BNN is trained for a set number of cycles to find the smallest number of descriptors possible via ARD. Therefore, the protocol used by the authors may not give an optimal BNN model. However, the current strategy within AstraZeneca is to obtain models that have a good balance between predictability and ease of interpretation (as few descriptors as possible) for the medicinal chemists. However, the maximum evidence method may give better predictability compared to the approach used here by the authors.

4.6. Evaluation of Model Predictivity. A number of statistical measures may be used to evaluate the models. We have used the square of the correlation coefficient to the y (measured pIC_{50}) = x (predicted pIC_{50}) line (r^2) and the root-mean-square error (rmse) as a measure of model performance. However, the r^2 obtained for the training set may not be reflective of the model performance on an independent test set.

A more realistic measure of model performance is based on how a model performs on compounds that it has not been trained on: the so-called test set compounds. The majority of the compounds in the test set of 249 oral drugs were reported as out of range data (no apparent inhibition at 2 mM); this affected 181 compounds. For the 68 compounds which have a pIC_{50} value, we have used the rmse and the square of the correlation coefficient in both the $y = x$ line (r^2) and the line of best-fit (r^2_{bf}) as a way of assessing the performance of each model.

Although the model is based on continuous pIC_{50} data, the model predictions have also been used in a classification sense because so many compounds in the test set were reported as out of range data. For classification, the 2 classes used are $pIC_{50} < 5$ and $pIC_{50} > 5$. We have used a $pIC_{50} = 5$ as a pragmatic cutoff for CYP1A2 inhibition. Compounds reported with $pIC_{50} > 5$ have potential to be problematic. Classification performance in a 2-class case is normally summarized in a confusion or error matrix that cross tabulates the observed and predicted patterns (Figure 4). In this paper we define (+) and (−) as $pIC_{50} > 5$ and $pIC_{50} < 5$, respectively. The total number of observations (N) in the test set is equal to $a + b + c + d$. For a review of methods for the assessment of prediction errors in classification, see Fielding and Bell.[50] A variety of error or accuracy measures can be calculated from a confusion matrix (Table 7). There are several measures that describe the accuracy of a single class (per class accuracy). The false positive rate is the proportion of negative cases that were incorrectly classified as positive, whereas the false negative rate is the proportion of positive cases that were incorrectly classified as negative. Sensitivity is the proportion of positive cases that were correctly predicted, and specificity is the proportion of negative cases that were correctly predicted. The positive and negative predictive power are the proportion of positive and negative predictions that were observed positive and negative, respectively. A naive measure of overall accuracy (all classes) is the correct classification rate. The problem with this measure is that it does not account for the fact that some of the apparent classification accuracy could be due to chance. For example, if a test set is 90% prevalent in one class, then it is possible to achieve an 82% correct classification rate by chance. The κ index of overall agreement for classification was developed by Cohen[51,52] and associates[53] in the context of psychology and psychiatric diagnosis. The κ index (eq 4) is considered to be superior to using correct classification rate as it assesses the model's improvement in prediction over chance. Landis and Koch[54] have suggested ranges of agreement

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{\text{total observed} - \text{chance agreement}} \qquad (4)$$

for the κ statistic (see Table 5). The κ index of agreement has very recently been highlighted in the field of QSAR to assess the overall accuracy of a model's predictive power.[55,56] However, as far as the authors are aware, this is the first comprehensive use of the κ index in the field of QSAR.

Table 7. Accuracy Measures That Can Be Calculated from a Confusion Matrix

measure	calculation
prevalence	$(a + c)/N$
correct classification rate	$(a + d)/N$
sensitivity	$a/(a + c)$
specificity	$d/(b + d)$
false positive rate	$b/(b + d)$
false negative rate	$c/(a + c)$
positive predictive power	$a/(a + b)$
negative predictive power	$d/(c + d)$
κ	$\{(a + d) - [((a + c)(a + b) + (b + d)(c + d))/N]\}/\{N - [((a + c)(a + b) + (b + d)(c + d))/N]\}$

Supporting Information Available: Training set data. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Brosen, K. Drug interactions and the cytochrome P450 system. The role of cytochrome P450 1A2. *Clin. Pharmacokinet.* **1995**, *29*, 20−25.

(2) Guengerich, F. P.; Shimada, T. Oxidation of toxic and carcinogenic chemicals by human cytochrome P-450 enzymes. *Chem. Res. Toxicol.* **1991**, *4*, 391−407.

(3) Jeppesen, U.; Loft, S.; Poulsen, H. E.; Broesen, K. A fluvoxamine-caffeine interaction study. *Pharmacogenetics* **1996**, *6*, 213−222.

(4) Shimada, T.; Iwasaki, M.; Martin, M. V.; Guengerich, F. P. Human liver microsomal cytochrome P-450 enzymes involved in the bioactivation of procarcinogens detected by umu gene response in Salmonella typhimurium TA 1535/pSK1002. *Cancer Res.* **1989**, *49*, 3218−3228.

(5) Yamashita, K.; Umemoto, A.; Grivas, S.; Kato, S.; Sato, S.; Sugimura, T. Heterocyclic amine-DNA adducts analyzed by 32P-postlabeling method. *Nucleic Acids Symp. Ser.* **1988**, *19*, 111−114.

(6) Zhai, S.; Dai, R.; Wei, X.; Friedman, F. K.; Vestal, R. E. Inhibition of methoxyresorufin demethylase activity by flavonoids in human liver microsomes. *Life Sci.* **1998**, *63*, 119−123.

(7) So, F. V.; Guthrie, N.; Chambers, A. F.; Moussa, M.; Carroll, K. K. Inhibition of human breast cancer cell proliferation and delay of mammary tumorigenesis by flavonoids and citrus juices. *Nutr. Cancer* **1996**, *26*, 167−181.

(8) Wattenberg, L. W. Inhibition of carcinogenesis by minor dietary constituents. *Cancer Res.* **1992**, *52*, 2085−2091.

(9) Steinmetz, K. A.; Potter, J. D. Vegetables, fruit, and cancer. II. Mechanisms. *Cancer, Causes Control* **1991**, *2*, 427−442.

(10) Moon, T.; Chi, M. H.; Kim, D.-H.; Yoon, C. N.; Choi, Y.-S. Quantitative Structure−Activity Relationships (QSAR) Study of Flavonoid Derivatives for Inhibition of Cytochrome P450 1A2. *Quant. Struct.-Act. Relat.* **2000**, *19*, 257−263.

(11) Höskuldsson, A. *Prediction Methods in Science and Technology*; Thor Publishing: Copenhagen, Denmark, 1996.

(12) Wold, S.; Albano, C.; Dunn, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. Multivariate Data Analysis in Chemistry. In *Chemometrics: Mathematics and statistics in Chemistry*; Kowalski, B. R., Ed.; D. Reidel Publishing Company: Dordrecht, Holland, 1984.

(13) Wold, S.; Eriksson, L.; Sjöström, M. *PLS in Chemistry, Encyclopedia of Computational Chemistry*; Wiley: 2000.

(14) Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Chapman and Hall/CRC and Place: Pacific Grove, CA, 1984.

(15) Steinberg, D.; Colla, P. *CART: Tree-Structured Non-Parametric Data Analysis*; Salford Systems: San Diego, CA, 1995.

(16) Neal, R. M. *Bayesian Learning for Neural Networks*; Lecture Notes in Statistics No. 118; Springer-Verlag: New York, 1996.

(17) Asikainen, A. H.; Ruuskanen, J.; Tuppurainen, K. A. Performance of (consensus) kNN QSAR for predicting estrogenic activity in a large diverse set of organic compounds. *SAR QSAR Environ. Res*. **2004**, *15*, 19−32.

(18) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. Two-Dimensional QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 276−285.

(19) Stouch, T. R.; Kenyon, J. R.; Johnson, S. R.; Chen, X.-Q.; Doweyko, A.; Li, Y. J. In silico ADME/Tox: why models fail. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 83−92.

(20) Bruneau, P. Search for Predictive Generic Model of Aqueous Solublity using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci*, **2001**, *41*, 1605−1616.

(21) Lewis, D. F. V. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics. *Biochem. Pharmacol*. **2000**, *60*, 293−306.

(22) Xu, Y.; Gao, H. Dimension related distance and its application in QSAR/QSPR model error estimation. *QSAR Comb. Sci.* **2003**, *22*, 422−429.

(23) McGinnity, D. F.; Griffin, S. J.; Moody, G. C.; Voice, M.; Hanlon, S.; Friedburg, T.; Riley, R. J. Rapid characterisation of the major drug-metabolising human hepatic cytochrome P-450 enzymes expressed in Escherichia coli. *Drug Metab. Dispos*. **1999**, *27*, 1017−1023.

(24) McGinnity, D. F.; Parker, A. J.; Soars, M.; Riley, R. J. Automated definition of the enzymology of drug oxidation by the major human drug metabolising cytochrome P450s. *Drug Metab. Dispos.* **2000**, *28*, 1327−1334.

(25) Obermeier, M. T.; White, R. E.; Yang, C. S. Effects of bioflavonoids on hepatic P450 activities. *Xenobiotica* **1995**, *25*, 575−584.

(26) Shimada, T.; Yamazaki, H.; Foroozesh, M.; Hopkins, N. E.; Alworth, W. L.; Guengerich, F. P. Selectivity of polycyclic inhibitors for human cytochrome P450s 1A1, 1A2, and 1B1. *Chem. Res. Toxicol.* **1998**, *11*, 1048−1056.

(27) Obach, R. S. Inhibition of human cytochrome P450 enzymes by constituents of St. John's Wort, an herbal preparation used in the treatment of depression. *J. Pharmacol. Exp. Ther*. **2000**, *294*, 88−95.

(28) Fuhr, U.; Wolff, S.; Harder, S.; Schymanski, P.; Staib, A. H. Quinoline inhibition of cytochrome P-450-dependent caffeine metabolism in human liver microsomes. *Drug Metab. Dispos*. **1990**, *18*, 1005−1010.

(29) Brosen, K.; Naranjo, C. A. Review of pharmacokinetic and pharmacodynamic interaction studies with citalopram. *Eur. Neuropsychopharmacol.* **2001**, *11*, 275−283.

(30) Von Moltke, L. L.; Greenblatt, D. J.; Duan, S. X.; Schmider, J.; Kudchadker, L.; Fogelmann, S. M.; Harmatz, J. S.; Shader, R. I. Phenacetin O-deethylation by human liver microsomes in vitro. Inhibition by chemical probes, SSRI antidepressants, nefazodone and venlafaxine. *Psychopharmacology* **1996**, *128*, 398−407.

(31) Kunze, K. L.; Wienkers, L. C.; Thummel, K. E.; Trager, W. F. Warfarin-Fluconazole. I. Inhibition of the human cytochrome P450-dependent metabolism of warfarin by fluconazole: in vitro studies. *Drug Metab. Dispos.* **1996**, *24*, 414−21.

(32) Langoue, S.; Furge, L. L.; Kerriguy, N.; Nakamura, K.; Guillouzo, A.; Guengerich, F. P. Mechanism-based inactivation of human cytochrome P-450 1A2 by oltipraz. *Abstracts of Papers*, 219th National Meeting of the American Chemical Society, San Francisco, CA, March 26−30, 2000; American Chemical Society: Washington, DC, 2000; TOXI-086.

(33) Kinzig-Schippers, M.; Fuhr, U.; Zaigler, M.; Dammeyer, J.; Rusing, G.; Labedzki, A.; Bulitta, J.; Sorgel, F. Interaction of pefloxacin and enoxacin with the human cytochrome P450 enzyme CYP1A2. *Clin. Pharmacol. Ther.* **1999**, *65*, 262−274.

(34) Shader, R. I.; Granda, B. W.; Von Moltke, L. L.; Giancarlo, G. M.; Greenblatt, D. J. Inhibition of human cytochrome P450 isoforms in vitro by zafirlukast. *Biopharm. Drug Dispos.* **1999**, *20*, 385−388.

(35) Katritzky, A.; Wang, Y.; Sild, S.; Tamm, T.; Kalrelson, M. QSPR Studies on Vapor Pressure, Aqueous Solubility, and the Prediction of Water-Air Partition Coefficient. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 720−725.

(36) Selma is an AstraZeneca in-house software package. For further information contact: Olsson, T.; Sherbukhin, V. Synthesis and Structure Administration (SaSA), AstraZeneca R&D Mölndal.

(37) *Physicians' Desk Reference*, 53rd ed.; Medical Economics Company: Montvale, 1999.

(38) Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *J. Med. Chem.* **2003**, *46*, 1250−1256.

(39) Jackson, J. E. *A User's Guide to Principal Components*; ISBN 0-471-62267-2; John Wiley: New York, 1991.

(40) Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multiway Principal Components and PLS-Analysis. *J. Chemom.* **1987**, *1*, 41−56.

(41) Baroni, M.; Costatino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D-QSAR Problems. *Quant. Struct.-Act. Relat*. **1993**, *12*, 9−20.

(42) http://www.JMPdiscovery.com (June 2002).

(43) Sarle, W. ftp://ftp.sas.com/pub/neural/FAQ3.html (February 2003).

(44) Ajay, W.; Murcko, M. Can We Learn To Distinguish Between "Drug-like" and "Nondrug-like" Molecule? *J. Med. Chem*. **1998**, *41*, 3314−3324.

(45) Burden, F.; Winkler, D. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem*. **1999**, *42*, 3183−3187.

(46) Burden, F.; Winkler, D. New QSAR Methods Applied to Structure-Activity Mapping and Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 236−242.

(47) Sorich, M. J.; McKinnon, R. A.; Miners, J. O.; Winkler, D. A.; Smith, P. A. Rapid Prediction of Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms Using Quantum Chemical Descriptors Derived with the Electronegativity Equalization Method. *J. Med. Chem.* **2004**, *47*, 5311−5317.

(48) Neal, R. M. Software for Flexible Bayesian Modeling. http://www.cs.utoronto.ca/∼radford (February 2003).

(49) Burden, F.; Ford, M.; Whitley, D.; Winkler, D. Use of Automatic Relevance Determination in QSAR Studies Using Bayesian Neural Networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1423−1430.

(50) Fielding, A. H.; Bell, J. F. A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ. Conserv.* **1997**, *24*, 38−49.

(51) Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37−46.

(52) Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* **1968**, *70*, 426−443.

(53) Fleiss, J. L.; Cohen, J.; Everitt, B. S. Large sample standard errors of Kappa and weighted Kappa. *Psychol. Bull.* **1969**, *72*, 323−327.

(54) Landis, J. R.; Koch, G. C. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159−174.

(55) O'Brien, S. E.; de Groot, M. J. Greater Than the Sum of Its Parts: Combining Models for Useful ADMET Prediction *J. Med. Chem.* **2005**, *48*, 1287−1291.

(56) Refsgaard, H. H. F.; Jensen, B. F.; Brockhoff, P. B.; Padkjær, S. B.; Guldbrandt, M.; Christensen, M. C. In Silico Prediction of Membrane Permeability from Calculated Molecular Parameters *J. Med. Chem.* **2005**, *48*, 805−811.

JM048959A