

Research Paper

Development of a New Predictive Model for Interactions with Human Cytochrome P450 2A6 Using Pharmacophore Ensemble/Support Vector Machine (PhE/SVM) Approach

Max K. Leong,^{1,2} Yen-Ming Chen,¹ Hong-Bin Chen,¹ and Po-Hong Chen¹

Received October 19, 2008; accepted December 8, 2008; published online December 23, 2008

Purpose. The objective of this investigation was to yield a generalized *in silico* model to quantitatively predict CYP2A6-substrates/inhibitors interactions to facilitate drug discovery.

Methods. The newly invented pharmacophore ensemble/support vector machine (PhE/SVM) scheme was employed to generate the prediction model based on the data compiled from the literature.

Results. The predictions by the PhE/SVM model are in good agreement with the experimental observations for those molecules in the training set ($n=24$, $r^2=0.94$, $q^2=0.85$, RMSE=0.30) and the test set ($n=9$, $r^2=0.96$, RMSE=0.29). In addition, this *in silico* model performed equally well for those molecules in the external validation sets, namely one set of benzene and naphthalene derivatives ($n=45$, $r^2=0.81$, RMSE=0.46) and one set of amine neurotransmitters ($n=4$, $r^2=0.98$, RMSE=0.32). Furthermore, when compared with crystal structures, the calculated results are consistent with the published CYP2A6-substrate co-complex structure and the plasticity nature of CYP2A6 is also revealed.

Conclusions. This PhE/SVM model is an accurate and robust model and can be utilized for predicting interactions with CYP2A6, high-throughput screening and data mining to facilitate drug discovery.

KEY WORDS: CYP2A6; IC₅₀; pharmacophore ensemble; plasticity; support vector machine.

INTRODUCTION

The polymorphic cytochrome P450 enzymes (CYPs) are best known for their role in the metabolism of a wide range of endogenous and xenobiotic molecules, including anticancer drugs and a variety of procarcinogens and promutagens (1–9). Inhibition of a single enzyme by co-administered multiple drugs, *viz.* polypharmacy (10), can substantially alter the plasma concentration of another drug, leading to adverse drug–drug interactions and undesired drug toxicity (11). Of all human CYP450 isozymes, CYP1A1, CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2D6, CYP2E1, CYP3A4 and CYP3A5 are involved in oxidation of more than 90% of environmental toxicants, drugs and carcinogens (12).

CYP2A6, which constitutes 5–10% of the total CYP in human liver (13), can metabolize some marketed drugs (14,15). Tegafur, for example, is metabolized to 5-FU by CYP2A6 to exert its anticancer efficacy (16,17). Furthermore, CYP2A6 is the primary enzyme responsible for metabolizing nicotine to its inactive metabolite cotinine (18), making CYP2A6 a putative smoking cessation treatment target by inhibiting nicotine metabolism (19,20). In addition, CYP2A6

can catalyze the 7-hydroxylation of coumarin (21), which is a toxic chemical compound found in many plants. Recent evidences also suggest the involvement of CYP2A6 in developing various types of cancer (22,23). Therefore, it is of critical importance to develop an *in silico* model to predict the interactions with CYP2A6 in the process of drug discovery in the hope of reducing the attrition rates due to adverse side effects as well as identifying inhibitors for smoke cessation and chemoprevention of CYP2A6-associated cancers.

A number of CoMFA 3D-QSAR models have been proposed (24–27) in addition to homology models (28–30) and a docking study (31). A brief summary can be found elsewhere (32). Nevertheless, the effects of protein flexibility, which are a pivotal factor to precisely determine the protein–ligand interactions, are ignored by these proposed models (33–35). In fact, pharmacophore or CoMFA 3D-QSAR modeling usually assumes that the target protein is rigid or with limited flexibility. Such premise is valid only when the target protein is relatively rigid. In other word, it undergoes very limited conformation changes when interacting with a variety of inhibitors or substrates, which usually is a fallible assumption when applied to many P450 enzymes (36) since it has been suggested that P450s can adopt multiple conformations upon binding with inhibitors or substrates (37). For example, recent published CYP3A4–ligand co-complex crystal structures indicate that the active site volume can increase by *ca.* 75% and 110% upon binding with ketoconazole and erythromycin, respectively (38), suggesting pronounced ligand-induced conformation changes. Even when binding with ligands

¹ Department of Chemistry, National Dong Hwa University, Shoufeng, Hualien 97401, Taiwan.

² To whom correspondence should be addressed. (e-mail: leong@mail.ndhu.edu.tw)

ABBREVIATIONS: AD, Application domain; PhE, Pharmacophore ensemble; SVM, Support vector machine.

of different sizes and shapes, the target protein may markedly change its active site volume and conformation as manifested by the CYP2B4–ligand co-complex structures (39) and the CYP2C5–substrate crystal structures (40). As a result, any proposed CYP2A6 analog-based models can be only applied to a very specific chemotype, which corresponds to a specific protein conformation when interacting with those ligands within the application domain (AD) of model generation.

In contrast to analog-based modeling, structure-based modeling seems to be a better alternative. If the target protein can adopt distinct conformations to interact with various substrate that is common in case of P450s (*vide supra*), an ensemble of protein conformations seemingly provide a more realistic approach (41). Recently, a number of crystal structures of CYP2A6 have been published, namely CYP2A6 in complexes with substrate coumarin (PDB code: 1Z10) as well as inhibitors methoxsalen (PDB code: 1Z11) (42), *N,N*-dimethyl(5-(pyridin-3-Yl)furan-2-Yl)methanamine (PDB code: 2FDU), *N*-methyl(5-(pyridin-3-Yl)furan-2-Yl)methanamine (PDB code: 2FDV), (5-(pyridin-3-Yl)furan-2-Yl)methanamine (PDB code: 2FDW) and 4,4'-dipyridyl disulfide (DPD) (PDB code: 2FDY) (43), of which it can be found that CYP2A6 can undergo substantial conformational change from the coumarin bound conformation to the DPD bound conformation as illustrated by Figure 3 of the publication by Yano *et al.* (43). In addition, the active site volume of the CYP2A6-coumarin co-complex is about 327 Å³ according to the estimation by the *CASTp* package (44) using a 1.4-Å probe, whereas that of the CYP2A6-DPD co-complex is about 470 Å³ or a 44% increase in size. As a result, it can be asserted that CYP2A6 can adopt more distinct conformations to interact with a variety of ligands. To accommodate the plasticity of CYP2A6 for more accurate predictions of interactions with CYP2A6, lengthy molecular dynamics (MD) calculations should be carried out. Nevertheless, CYP2A6 and any other CYP450 isozymes are heme proteins, *viz.* transitional-metal-containing systems, which can only be modeled by quantum mechanics or QM/MM methods (45) that has been further explained in detail elsewhere in the case of CYP450s (46). Consequently, such computationally expensive QM or QM/MM MD calculations make such protein conformation sampling extremely difficult in some cases, if not absolutely infeasible.

As a result, any analog- or structure-based modeling methods that fail to take into account protein plasticity will give rise to fallible predictions or substantial deviations when the target protein is highly flexible. Recently, a novel scheme has been proposed, in which a panel of plausible pharmacophore hypothesis candidates were assembled to construct a pharmacophore ensemble (PhE), which, in turn, was treated as input for regression analysis via support vector machines (SVM) (47). Each pharmacophore member in the PhE represents a protein conformation or a number of protein conformations with closed spatial arrangements. Unlike any other analog-based modeling methods, this PhE/SVM scheme can take into account protein plasticity, which is of critical importance to be addressed when the target protein can adopt significantly various conformations to interact structurally diverse ligands (48), by using PhE in place of protein conformation ensemble. More importantly, this PhE/SVM has been applied to study the liability of human *ether-a-go-go-*

related gene (47) and CYP2B6-substrate interactions (49), both of which are highly flexible proteins. The aim of this study was to derive an *in silico* model based on PhE/SVM scheme to predict interactions with human CYP2A6.

MATERIALS AND METHODS

Data Compilation

Data enlisted in this investigation were compiled from different literature sources (50–54). IC₅₀ values were taken for those compounds, which were measured by inhibition of coumarin 7-hydroxylation since it is considered to be well characterized (55) and, most importantly, it provided the largest quantity of consistent data records. Furthermore, chemical structures were cautiously examined and only compounds with defined stereochemistry were assembled. All molecules enrolled in this study, their corresponding biological activities and references to the literature are listed in Table I.

Table I. Selected Compounds for this Study, Their IC₅₀ (μM) Values or Average Values if Applicable and References

Molecules	IC ₅₀ (μM)	Refs
β-Nicotyrine	2.20	(53)
4-Hydroxy-1-(3-pyridyl)-1-butanone	4.20	(50)
2-(<i>p</i> -Tolyl)-ethylamine	4.90	(52)
4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone	5.00	(50)
4-Oxo-1-(3-pyridyl)-1-butanone	6.80	(50)
4-Methoxybenzaldehyde	7.10	(52)
4-Methylbenzaldehyde	13.00	(52)
(–)-Menthone oxime	24.61	(54)
7-Methylcoumarin	30.00	(51)
(+)-Menthol	37.77	(54)
2-Phenylethylamine	39.00	(52)
Butylcyclohexane	43.00	(51)
Indan	50.00	(51)
(+)-Neomenthol	52.77	(54)
2-Indanone	57.00	(51)
Thymol	67.49	(54)
(–)-Menthone	67.54	(54)
(–)-Menthol	70.49	(54)
Butylbenzene	75.00	(51)
Biphenyl	82.00	(51)
Benzaldehyde	120.00	(52)
(<i>R</i>)-(+)-pulegone	129.50	(54)
(<i>S</i>)-(–)-pulegone	139.40	(54)
2,3-Dihydrobenzofuran	180.00	(51)
2-Coumarone	300.00	(51)
2-Benzoxazolinone	870.00	(51)
4,6-Dimethyl-α-pyrone	1,600.00	(51)
ε-Caprolactone	16,000.00	(51)
4-Methoxy-2(5H)-furanone	18,000.00	(51)
2H-pyran-2-one	22,000.00	(51)
σ-Valerolactone	29,000.00	(51)
5,6-Dihydro-2H-pyran-2-one	33,000.00	(51)
γ-Butyrolactone	350,000.00	(51)

Table III. Experimentally Observed pIC₅₀ Values of Compounds in the Test Set, Corresponding Predicted Values by Hypo A, Hypo B, Hypo C and SVM Model, Residual (Δ) and Associated Statistic Numbers (Correlation Coefficient) r^2 , RMSE, Maximum Residual, Average Residual and Standard Deviation of Residual

Molecules	Obs.	Hypo A		Hypo B		Hypo C		SVM model	
	pIC ₅₀	pIC ₅₀	Δ	pIC ₅₀	Δ	pIC ₅₀	Δ	pIC ₅₀	Δ
4-Hydroxy-1-(3-pyridyl)-1-butanone	5.38	6.00	0.62	5.34	-0.04	5.92	0.54	5.51	0.13
2-(<i>p</i> -tolyl)-ethylamine	5.31	4.92	-0.39	5.12	-0.19	5.60	0.29	5.43	0.12
4-Methoxybenzaldehyde	5.15	5.07	-0.08	5.70	0.55	4.92	-0.23	5.43	0.28
2-Phenylethylamine	4.41	5.10	0.69	4.49	0.08	5.37	0.96	4.84	0.43
<i>R</i> -Neomenthol	4.28	3.80	-0.48	3.47	-0.81	3.80	-0.48	3.72	-0.56
2,3-Dihydrobenzofuran	3.74	4.82	1.08	3.19	-0.55	4.85	1.11	3.77	0.03
2-Coumarone	3.52	3.62	0.10	4.64	1.12	3.30	-0.22	3.90	0.38
ϵ -Caprolactone	1.80	1.92	0.12	1.74	-0.06	2.10	0.30	1.80	0.00
σ -Valerolactone	1.54	1.72	0.18	1.96	0.42	1.96	0.42	1.69	0.15
r^2		0.88		0.84		0.87		0.96	
RMSE			0.52		0.55		0.59		0.29
Max			1.08		1.12		1.11		0.56
Average			0.42		0.42		0.51		0.23
SD			0.34		0.37		0.32		0.19

corresponding IC₅₀ values. The *HypoGen* module was used to develop chemical feature-based pharmacophore hypotheses using hydrogen-bond acceptor (HBA), hydrogen-bond acceptor lipid (HBA lipid), hydrogen bond donor (HBD), hydrophobic (HP) and ring aromatic (RA) as feature candidates. The minimum and maximum numbers of each selected chemical feature and total features were varied in order to find better performance. In addition, to maximize the hypothesis diversity, assorted combinations of variable weight and variable tolerance were employed. Both fitting algorithms, namely the “best” fit and the “fast” fit, were also tested. The generated pharmacophore models were then used to predict the IC₅₀ values of those compounds in the test set.

The cost of a generated hypothesis and that of its associated null hypothesis were taken from the log file, and the difference between these two values was calculated to survey the statistic quality of a hypothesis. In addition, statistical analyses, namely the correlation coefficient (r^2), root-mean-square error (RMSE), maximum residual, average residual and standard deviation of residual between the observed and predicted IC₅₀ values, were computed for both training set and test set. Only those pharmacophore models that showed good prediction accuracy and excellent statistic performance in both sets were eligible to construct the PhE.

SVM Calculations

The predicted pIC₅₀ values of those compounds in the training set by those pharmacophore hypotheses in the PhE were treated as input for regression calculations using the *LIBSVM* package (60), which consists of two routines for regression, namely *svm-train* and *svm-predict*, to develop an SVM model, based on the input data and options, and to predict the test samples using a model previously built with *svm-train*, respectively. Two regression modes, namely ϵ -SVR and ν -SVR, were also tested. The frequently used kernel radial basis function (RBF) was used for its simplicity and prominent performance (61). The SVR models were gener-

ated based on various runtime parameters, which were carried out using an in-house perl script to systematically scan through those runtime parameters, namely cost C , the width of the RBF kernel γ and ϵ and ν in cases of ϵ -SVR and ν -SVR, respectively. In addition, the generated SVM models were further validated by a 10-fold cross-validation instead of the most popular leave-one-out one since the former has been proven to perform better than the latter (62).

External Validation

It can be argued that a high value of cross-validated q^2 does not suffice to warrant the predictivity of a theoretical model (63). It may be necessary to further evaluate the real performance of a prediction model by an external test set, which is not involved in model development and, consequently, exerts no effect on the prediction model. As a result, it is highly possible that a prediction model is statistically authentic if the model has high values of q^2 and r^2 calculated by the cross-validation and external validation, respectively.

Therefore, the generated PhE/SVM model was further challenged with a group of benzene and naphthalene derivatives and a series of neurotransmitters and steroids published by Rahnasto *et al.* (26) and Higashi *et al.* (64), respectively. These molecules were selected as external test sets 1 and 2, respectively, and the experimental IC₅₀ values measured by the inhibition of coumarin 7-hydroxylation were adopted to be consistent with the compound selections in the training set and the test set.

Comparisons with Crystal Structures

To further elucidate the authenticity of constructed models, a number of molecules were selected to compare with recently published crystal structures (42,43) (*vide supra*). Those selected molecules were first mapped onto the pharmacophore hypotheses and the matched conformations were then extracted and rigidly aligned with the ligands in the

Table IV. Costs of Returned Hypotheses and Null Hypotheses and the Cost Differences (Δ) Between Returned and Null Hypotheses for the Pharmacophore Models Hypo A, Hypo B, and Hypo C

Cost	Hypo A	Hypo B	Hypo C
Null hypothesis	282.79	282.79	282.79
Returned hypothesis	145.20	155.53	150.59
Δ	137.59	127.26	132.20

co-complex structures. Finally, the aligned structures along with the associated pharmacophore models were placed into the corresponding co-complex structures.

RESULTS AND DISCUSSION

PhE

Of all generated pharmacophore hypotheses using different combinations of chemical features and runtime conditions, three models, designated by Hypo A, Hypo B and Hypo C, were enlisted to construct the PhE based on the prediction accuracy of individual molecule and the statistical analyses in the training set and the test set as well as the cost differences as shown in Tables II, III and IV, respectively. These three candidate models in the ensemble consist of the same chemical features, namely one HBA lipid, two HPs and one RA. Tables V, VI and VII summarize the characteristics of these three hypotheses, including weights, tolerances, three-dimensional coordinates and interfeature distances.

These three pharmacophore hypotheses are spatially arranged differently despite the fact that they possess the same chemical features as demonstrated by Fig. 1. The distance between two HP groups in Hypo A, for instance, is 6.274 Å, whereas that slightly decreases to 6.010 Å in Hypo B

and increases to 6.288 Å in Hypo C. The lengths between the chemical features HP and RA are 1.256, 1.465 and 1.472 Å in Hypo A, Hypo B and Hypo C, respectively. More pronounced variations in interfeature distance can be found from the distance between the chemical features HP and HBA lipid, which show the maximum by Hypo B with a value of 5.632 Å, followed by Hypo A with a value of 2.901 Å and the minimum by Hypo C with a value of 2.640 Å. The spatial discrepancies among these three models can also be illustrated by two angles centered at either one of HPs and connecting to the other HP and RA or HBA lipid, varying from 77.1° and 112.2° in Hypo A to 75.7° and 50.5° in Hypo B and 81.1° and 56.3° in Hypo C.

These three models in the PhE differ not only in the relative topological relationships but also in the absolute coordinates in the space as illustrated by the superposition of these three models illustrated in Fig. 2. As shown in Fig. 3, such differences can be further depicted by fitting these three models into 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK). Thus, NNK adopts different conformations to generate the best fit with these models. The chemical feature RA, for example, is mapped onto the pyridine moiety for all of these three models. Conversely, the chemical feature HBA lipid matches the nitroso functional group in Hypo A and Hypo C, whereas it coincides with the carbonyl functional group in Hypo B. The observation of such discrepancy becomes even more pronounced by the overlay of these three conformations as demonstrated in part D of Fig. 3, suggesting that all of these models adopted various conformations or orientations to exert the biological activities and a PhE is needed to address the conformational plasticity as a result.

The maximum error of Hypo C in the training set was yielded from the prediction of butylcyclohexane with a residual of 1.63, whereas Hypo A and Hypo B gave rise to residuals of 0.52 and 1.16, respectively (Table II). The prediction of 4-oxo-1-(3-pyridyl)-1-butanone by Hypo A

Table V. Weights, Tolerances, Three-Dimensional Coordinates of Chemical Features and Interfeature Distances of Pharmacophore Model Hypo A



	HBA lipid		Hydrophobic		Ring aromatic	
Weights	1.48		2.12		4.02	
Tolerances	1.30	1.90	1.30	1.30	1.90	1.30
X	1.95	0.62	-0.91	-3.15	-3.86	-3.08
Y	0.46	-2.13	0.45	6.21	5.47	3.04
Z	-2.43	-1.73	-1.97	-0.82	-0.09	1.48
						
HBA lipid	3.0					
Hydrophobic	2.9	3.0				
Hydrophobic	7.8	9.2	6.3			
Ring aromatic	8.0	9.0	6.1	1.3		
	6.9	7.1	4.8	3.9	3.0	

Table VI. Weights, Tolerances, Three-dimensional Coordinates of Chemical Features and Interfeature Distances of Pharmacophore Model Hypo B

	HBA lipid		Hydrophobic	Hydrophobic	Ring aromatic	
Weights	1.63		2.33	3.03	4.42	
Tolerances	1.30	2.20	1.45	1.45	1.75	1.30
X	1.25	1.45	-2.66	-2.66	-1.29	-0.26
Y	2.96	5.75	-0.50	-0.50	0.01	-0.82
	0.01	1.11	0.44	0.44	0.34	-2.35
	○ →				○ →	
HBA lipid	3.0					
Hydrophobic	5.6	6.7				
Hydrophobic	5.0	7.7	6.0			
Ring aromatic	3.7	6.5	5.8	1.5		
	4.8	7.6	5.7	3.7	3.0	

deviated most from the observed value with a residual of 0.79, whereas Hypo B and Hypo C only showed deviations of 0.45 and 0.42, respectively. The largest deviation by Hypo B was resulted from the prediction of 2-benzoxazolinone with a value of 1.49, whose evaluation errors were merely 0.56 and 0.15 by Hypo A and Hypo C, respectively. Similarly, Hypo A perfectly predicted 4-methoxy-2(5H)-furanone with no error, Hypo B only generated a residual of 0.02, whereas Hypo C yielded a significant deviation of 0.90. In fact, such prediction discrepancies for molecules in the training set among these three models in the PhE render the fact that no single pharmacophore hypothesis performed better than the others

for all molecules in the training set; nor did one perform worse than the others. Generally, it can be found from Table II that these three hypotheses in the PhE gave rise to very close prediction trend, resulting in high values of correlation coefficient r^2 (Table II), and it can be further demonstrated by the scatter plot of observed vs. the predicted pIC_{50} values as illustrated in Fig. 4.

Furthermore, statistical parameters root-mean-square error, maximal residual, average residual and standard deviation of residuals in the training set (Table II) suggest that these three models in the PhE functioned equally well in the training set. More importantly, it is highly possible that

Table VII. Weights, Tolerances, Three-Dimensional Coordinates of Chemical Features and Interfeature Distances of Pharmacophore Model Hypo C

	HBA lipid		Hydrophobic	Hydrophobic	Ring aromatic	
Weights	1.64		1.64	2.34	4.44	
Tolerances	1.60	2.20	1.60	1.60	1.60	1.60
X	-1.69	-1.77	0.76	-1.40	-0.58	1.90
Y	-2.19	-4.34	-3.18	0.26	-0.70	0.98
Z	3.61	5.71	3.72	-1.08	-1.84	-2.07
	○ →				○ →	
HBA lipid	3.0					
Hydrophobic	2.6	3.4				
Hydrophobic	5.3	8.2	6.3			
Ring aromatic	5.8	8.5	6.2	1.5		
	7.4	10.1	7.2	3.5	3.0	

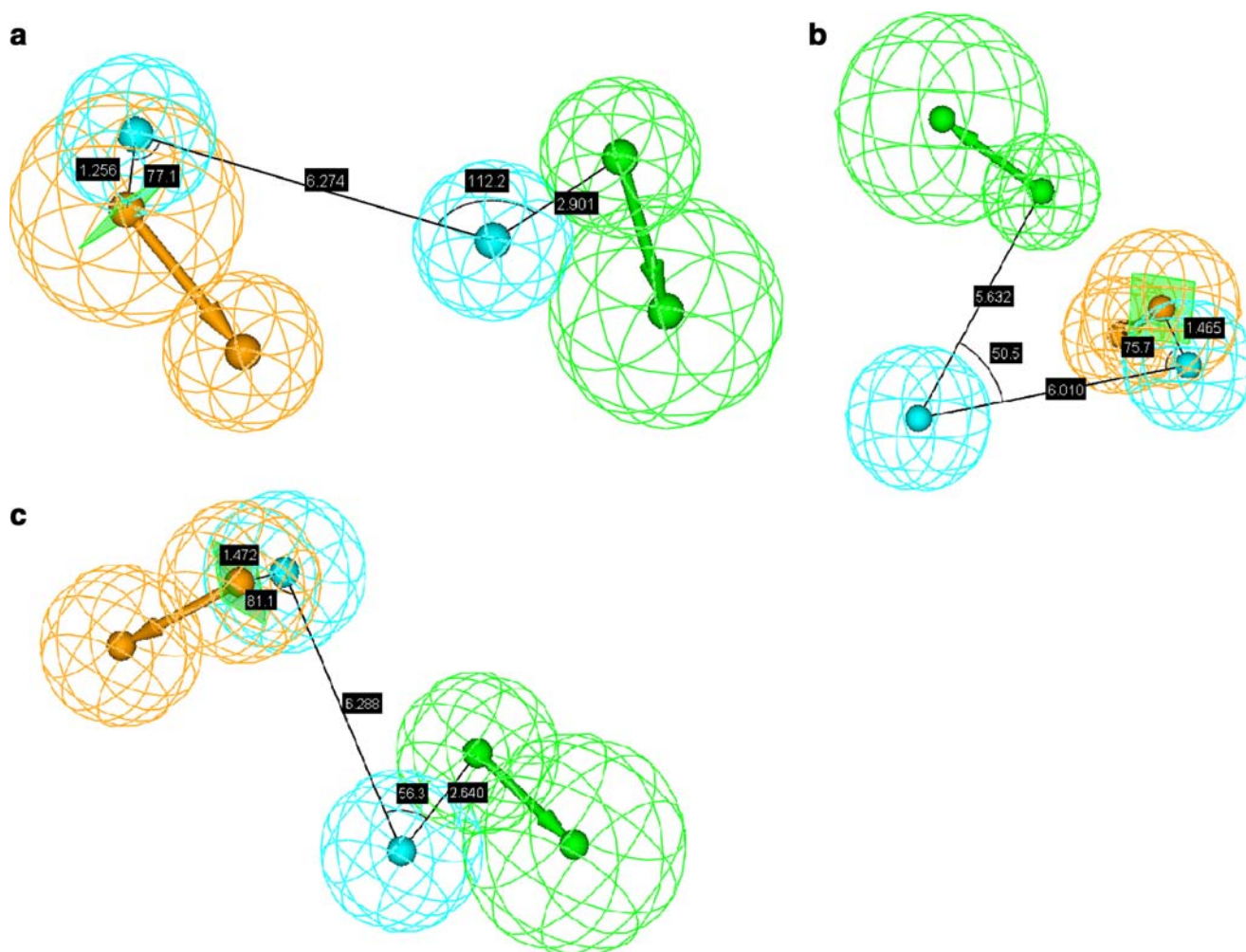


Fig. 1. Generated pharmacophore models **A** Hypo A, **B** Hypo B and **C** Hypo C, consisting of lipid hydrogen-bond acceptor (*green*), hydrophobic (*light blue*) and ring aromatic (*orange*) chemical features. The interfeature distances and angles among features, depicted in *white*, are measured in Ångstroms and degrees, respectively.

these three models are statistically authentic models since the cost differences between the null hypothesis and returned hypotheses are 137.59, 127.26 and 132.20 for Hypo A, Hypo B and Hypo C, respectively (Table IV), all of which are substantially larger than 60 that is the cost difference required to reach the level of a more than 90% chance to show the statistical correlation between the hypothesis and the input data as described in the *Catalyst's* manual.

Various levels of performance decline can be found when applied these three models in the PhE to predict those molecules in the test set in terms of the statistical evaluations, namely RMSE, average residual and residual standard deviation as shown in Table III. Nevertheless, the maximum deviations calculated by Hypo B and Hypo C slightly decreased from 1.49 and 1.63 in the training set to 1.12 and 1.11 in the test set, respectively. Prediction discrepancies among these three pharmacophore models found in the training set can also be found in the test set. The prediction of 2-coumarone by Hypo B, for example, resulted in the maximal error of 1.12, whereas Hypo A and Hypo C only generated errors of 0.10 and 0.22, respectively. The maximal

deviations by Hypo A and Hypo C were yielded from the predictions of 2,3-dihydrobenzofuran, which was only 0.55 produced by Hypo B. As a result, discrepancies in the prediction trends by these three models can also be found in the test set as displayed in Fig. 5.

In general, the predictions by Hypo A, Hypo B and Hypo C are, in general, in agreement with observed values for molecules in both the training set and the test set as shown in Tables II and III. In fact, they show very similar performance in both training set and test set by comparing their r^2 values (Tables II and III). The differences in the parameter r^2 calculated by Hypo A, Hypo B and Hypo C between the training set and the test set were only +0.01, -0.03 and +0.02, respectively. Such negligible differences in r^2 suggest that they were statistically well-trained models in contrast to an over-trained model, which otherwise will give rise to a considerable r^2 difference between both sets. Therefore, it can be asserted that Hypo A, Hypo B and Hypo C are qualified candidates for the PhE development based on their performances in the training set and the test set as well as their statistical evaluations as mentioned above.

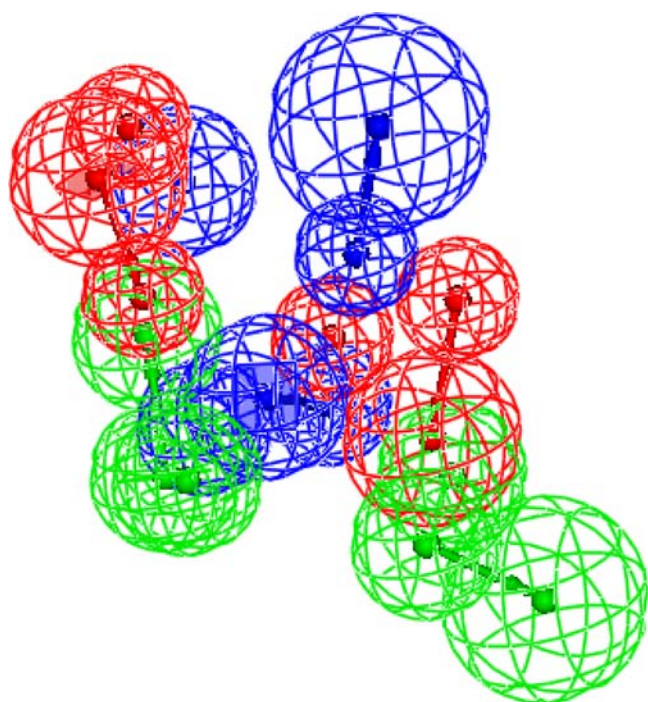


Fig. 2. Superposition of three pharmacophore models Hypo A, Hypo B and Hypo C, denoted in green, red and blue, respectively.

SVM

The optimal SVM, whose input parameters are summarized in Table VIII, was chosen from various runtime conditions based on the prediction calculations of those molecules in the training set and 10-fold cross-validation as given in Table II. The prediction results for those molecules in the test set are listed in Table III. It can be found from Table II that the SVM model yielded smaller residuals than the maximal deviations produced by those hypotheses in the PhE for any given molecule in the training set. The SVM model even gave rise to the smallest residuals in some cases. The prediction of 5,6-dihydro-2H-pyran-2-one by SVM, for instance, resulted in an error of 0.10, whereas Hypo A, Hypo B and Hypo C yielded deviations of 0.20, 0.22 and 0.44, respectively. As a result, most of points predicted by SVM generally lie on or are very closer to the regression line with slope of 1.00, *i.e.* the ideal regression line, as compared with Hypo A, Hypo B and Hypo C as shown in Fig. 4.

Furthermore, all of the statistical parameters, shown in Table II, support the fact that the SVM model performed better than any of pharmacophore models in the PhE in the training set except maximal residuals, which were 0.79 by Hypo A and 0.85 by SVM. Nevertheless, of 24 molecules in the training set, Hypo A produced 11 predictions, whose residuals were more than 0.50, whereas Hypo B, Hypo C and SVM only yielded 3, 5 and 2, respectively. The 10-fold cross-validation of the SVM model produced the correlation coefficient q^2 of 0.85 as compared with an r^2 of 0.94 for the training set as indicated in Table II. The inconsiderable discrepancy between both two parameters signifies the fact that the SVM model shows highly statistical significance between the theoretical model and the input data and, more

importantly, it is highly possible that this SVM model is an authentic model.

Unlike all models in the PhE, the SVM model showed various levels of performance improvement when applied to the test set as indicated by all statistical parameters (Tables II and III) except average residuals, which were 0.21 in the training set and 0.23 in the test set. Consequently, the SVM model performed better than any of pharmacophore model in the PhE in the test set as asserted by all of statistical estimations shown in Table III. In addition, the SVM model only gave rise to one prediction, which deviated from the experimental value by more than 0.50, in the test set, whereas Hypo A, Hypo B and Hypo C generated 3, 4 and 3, respectively, resulting in smaller distances from the ideal prediction line as illustrated in Fig. 5.

It is of practical importance to evaluate an *in silico* model by taking into consideration its performance in both the training set and the test set. As a result, it can be concluded that the SVM model outperformed Hypo A, Hypo B and Hypo C based on the prediction and statistical performances in both sets mentioned above presumably because of the fact that the PhE/SVM approach cannot only take into account the protein conformational flexibility and but also gives rise to the more realistic final model that, nevertheless, cannot be achieved by traditional ligand-based modeling schemes.

External Validation

The prediction results of those molecules in the external validation sets 1 and 2 as well as their associated statistical numbers are listed in Table IX. There were 48 molecules, whose inhibition activities of CYP2A6 were investigated by Rahnasto *et al.* (26), consisting of a variety of chemical structures, namely naphthalenes, quinolines, tetralones and non-planar compounds. Three molecules were excluded from the selection due to their uncertain biological activities. The remaining 45 molecules were selected to constitute the external validation set 1. It can be observed from Fig. 6, in which all of molecules in all sets were projected into the chemical space, spanned by three principal components, that some of molecules in the external validation set 1 are surrounded by molecules in the training set, whereas most of molecules in the external validation set 1 lie outside the boundary of the training set molecules, suggesting that some of molecules in the external validation set 1 are covered within the AD of model generation, whereas the others are located outside the AD. As a result, the predictions of those molecules within and outside the AD will be interpolation and extrapolation in nature, respectively. In addition, the biological activities, spanning over about 6 orders of magnitude, imply the diverse nature of these 45 molecules in the external validation set 1. Consequently, the structural and biological versatility renders the fact those molecules in the external validation set 1 serve good samples to verify the predictivity of the generated PhE/SVM model.

It can be observed from Table IX that the predictions by PhE/SVM are generally in good agreement with observed values for most of molecules in the external validation set 1 as manifested by the fact that, of those 45 molecules in the external validation set 1, there were 35 molecules, whose absolute residuals calculated by the PhE/SVM model were

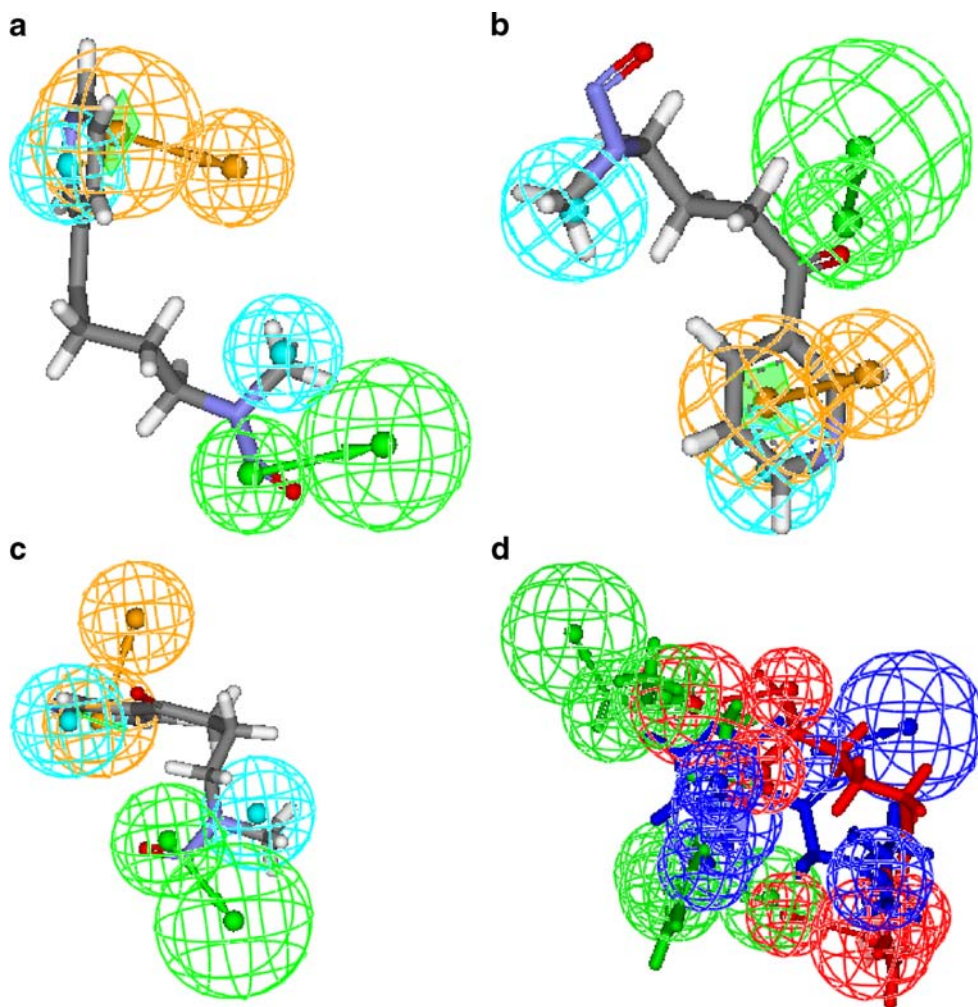


Fig. 3. Pharmacophore models **A** Hypo A, **B** Hypo B and **C** Hypo C fitted to 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone and **D** overlay of these three models, which are color-coded by *green, blue and red*. The chemical features are described in Fig. 1.

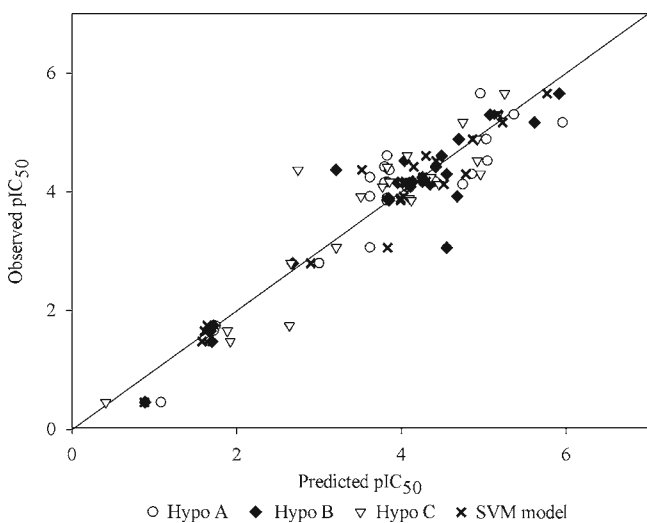


Fig. 4. Observed pIC_{50} vs. the pIC_{50} predicted by Hypo A, Hypo B, Hypo C and SVM model for those molecules in the training set and the ideal regression line.

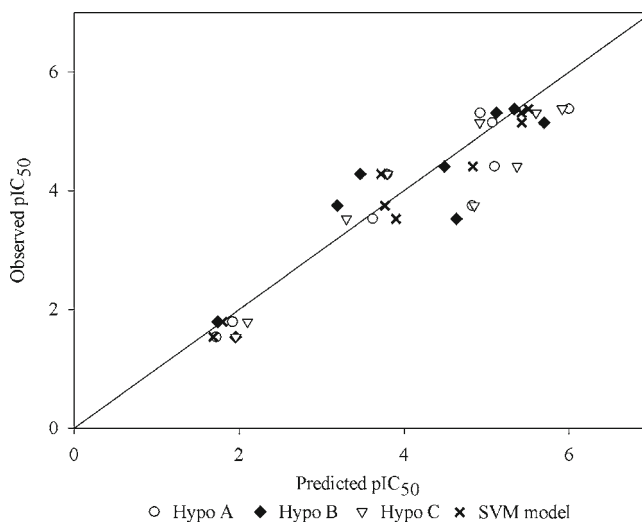


Fig. 5. Observed pIC_{50} vs. the pIC_{50} predicted by Hypo A, Hypo B, Hypo C and SVM model for those molecules in the test set and the ideal regression line.

Table VIII. Optimal Runtime Parameters for the SVM Model

Parameter	Value
SVM type	ϵ -SVR
Kernel type	Radial basis function
γ	0.001
Cost	100000
ϵ	0.1

less than 0.50. The maximum prediction error in the external validation set 1 was 1.47, yielded by the estimation of cotinine, which slightly increased from the training set, implying that the PhE/SVM performance slightly declined from the training set to the external validation set 1. Such deterioration can be plausibly attributed to the fact that most of the predictions were extrapolation in nature as mentioned above. Nevertheless, the insignificant differences in r^2 , RMSE, average residual and standard deviation of residual between both sets assert the fact that this PhE/SVM model can maintain similar level of predictivity when applied to the interpolation and extrapolation samples, which is of critical importance to a prediction model.

Of 18 compounds studied by Higashi *et al.* (64), histamine, serotonin, dopamine and tryptamine were selected as the external validation set 2 since they were only molecules, whose biological activities were well characterized. These molecules were more serious challenges to the generated PhE/SVM model than all of the molecules in the test set and the external validation set 1 since they are endogenous compounds *per se* in contrast to all molecules in the other sample sets designated in this study, which completely are xenobiotics. The dissimilarity of these 4 compounds in the external test set 2 from the others can be illustrated by Fig. 6, in which it can be found that the distances between those molecules in the external validation set 2 and any other molecules are considerably far, and, in addition, they are completely outside the AD of model generation, suggesting that those 4 endogenous molecules are outliers with respect to those xenobiotic molecules.

When applied to the external validation set 2, the PhE/SVM model performed extraordinarily well as shown in Table IX, giving rise to deviations from the experimental values within 0.54 log units and an average residual of 0.27. In addition, the predictions and experimental values were correlated extremely well with an r^2 value of 0.98. It can be asserted that the PhE/SVM model demonstrated excellent performance in the external validation set 2 based on the prediction accuracy as well as the statistical evaluations. More importantly, the PhE/SVM is very insensitive to the outliers, suggesting that it is very robust (65), which is an important fact to a prediction model.

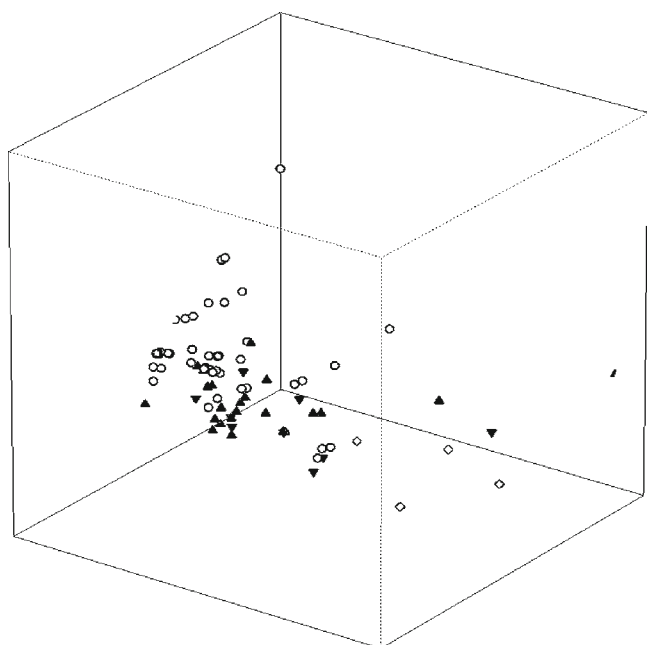
Thus, it is plausible to conclude, based on the performance in both external validation sets, that this *in silico* model based on the PhE/SVM scheme is an accurate and robust model to predict the interactions between CYP2A6 and substrates/inhibitors. Furthermore, this prediction model can be applied to a variety of chemical structures with high levels of prediction accuracy, which can be attributed to the fact that this PhE/SVM model can take into account the protein conformational flexibility while interacting with

Table IX. Experimentally Observed pIC₅₀ Values of Compounds in the External Validation Sets, Corresponding Predicted Values by the PhE/SVM Model, Residual (Δ) and Associated Statistic Numbers (Correlation Coefficient) r^2 , RMSE, Maximum Residual, Average Residual and Standard Deviation of Residual

Molecules	Obs.	Pred.	
	pIC ₅₀	pIC ₅₀	Δ
External validation set 1 ^a			
1,2-Dichloronaphthalene	4.82	5.04	0.22
1,2-Dimethylnaphthalene	4.59	4.69	0.10
1,3-Dimethylnaphthalene	5.08	5.11	0.03
1,4-Dichloronaphthalene	5.60	5.15	-0.45
1,5-Dichloronaphthalene	5.60	5.31	-0.29
1,5-Dimethylnaphthalene	4.51	4.73	0.22
1,6-Dimethylnaphthalene	3.89	4.45	0.56
1,7-Dimethylnaphthalene	4.51	4.58	0.07
1-Chloronaphthalene	4.74	4.73	-0.01
1-Methylisoquinoline	4.22	4.47	0.25
1-Methylnaphthalene	4.47	4.68	0.21
1-Naphthol	3.89	4.17	0.28
2,4-Dimethylquinoline	3.08	3.06	-0.02
2,6-Dimethylnaphthalene	5.00	5.57	0.57
2,6-Dimethylquinoline	3.55	3.72	0.17
2,7-Dimethylnaphthalene	5.77	6.22	0.45
2,7-Dimethylquinoline	3.40	4.58	1.18
2-Bromonaphthalene	6.26	6.48	0.22
2-Chlorobiphenyl	4.44	4.61	0.17
2-Chloronaphthalene	5.27	5.84	0.57
2-Ethyl-naphthalene	4.92	4.98	0.06
2-Fluoronaphthalene	6.17	4.99	-1.18
2-Methoxynaphthalene	4.21	4.21	0.00
2-Methylnaphthalene	5.62	6.16	0.54
2-Naphthol	3.85	3.99	0.13
3-Methylisoquinoline	4.60	4.74	0.13
3-Methylquinoline	3.70	4.18	0.48
4-Chlorobiphenyl	3.82	3.91	0.09
6,7-Dimethoxy-2-tetralone	2.70	2.73	0.03
7-Methyl-2-naphthaldehyde	5.17	5.55	0.38
Dibromo- <i>p</i> -xylo	4.23	4.37	0.14
Dichloro- <i>p</i> -xylo	4.77	4.47	-0.30
Naphthalene	4.60	4.71	0.11
Nicotine	3.24	3.83	0.59
Phenanthrene	4.01	4.19	0.18
Quinaldine	3.72	4.12	0.40
α -Tetralone	4.28	4.36	0.08
β -Tetralone	4.68	4.89	0.21
Cotinine	1.46	2.93	1.47
Benzaldehyde	3.92	4.01	0.09
4-Methylbenzaldehyde	4.88	4.87	-0.01
4-Methoxybenzaldehyde	5.15	5.53	0.38
amphetamine	3.50	4.39	0.89
2-(<i>p</i> -tolyl)-ethylamine	5.30	5.82	0.52
2-Phenylethylamine	4.41	4.56	0.15
r^2		0.81	
RMSE			0.46
Max			1.47
Average			0.32
SD			0.42
External validation set 2 ^b			
Histamine	3.22	3.46	0.24
Serotonin	3.40	3.94	0.54
Tryptamine	6.10	6.34	0.24
Dopamine	3.93	3.99	0.06
r^2		0.98	
RMSE			0.32
Max			0.54
Average			0.27
SD			0.20

^a Rahnasto *et al.* (26)

^b Higashi *et al.* (64)



▲ Training Set ▼ Test Set ○ External Set 1 ◇ External Set 2

Fig. 6. Molecular distribution for those samples in the training set (filled triangle), the test set (filled inverted triangle), the external validation set 1 (open circle) and the external validation set 2 (open diamond) in the chemical space spanned by three principal components.

structurally diverse small molecules that, in turn, is of critical importance and yet often neglected by most of the analogue-based modeling methods. Consequently, it is plausible to expect that this PhE/SVM model will show similar prediction performance when applied to other molecules of different chemotypes.

Comparisons with Crystal Structures

7-Methylcoumarin was chosen to compare with bound coumarin in the co-complex structures since, of 33 molecules enlisted in this study and 6 substrates in the published co-complex structures, 7-methylcoumarin and coumarin are the most similar molecules with only difference in methyl group. The mapped conformation of 7-methylcoumarin by Hypo C was employed since the prediction by Hypo C generated the smallest deviation from the observed value as compared with that by Hypo A and Hypo C. As a result, it is plausible to assume that 7-methylcoumarin will adopt the same conformation or a very similar one to this one to interact with CYP2A6 enzyme.

It can be observed from Fig. 7, which displays the overlay of 7-methylcoumarin and CYP2A6-coumarin co-complex, that 7-methylcoumarin was perfectly aligned with coumarin, which presumably can be attributed to the fact that both molecules are very rigid and flat, resulting in very limited conformational flexibility. The two hydrophobic chemical features found in 7-methylcoumarin presumably are due to the interactions with residues of Phe-107, Gly-301, Thr-305 and Phe-480, which also can be confirmed by the analysis using the *LigPlot* program (66) that those four residues form hydrophobic contacts with coumarin. The measured distance between carbonyl oxygen of 7-methylcoumarin and one of amine hydrogens of Asn-297 is 2.532 Å as illustrated in Fig. 7, suggesting the formation of a hydrogen bond between two moieties as observed in the crystal structure (42), which is completely consistent with the derived Hypo C.

It was proposed that the π - π stacking interaction took place between the aromatic ring of Phe-107 and that of coumarin (42). Nevertheless, the estimated distance between the aromatic ring of 7-methylcoumarin and that of Phe-107 is 7.16 Å, whereas that is only 6.51 Å between the aromatic ring

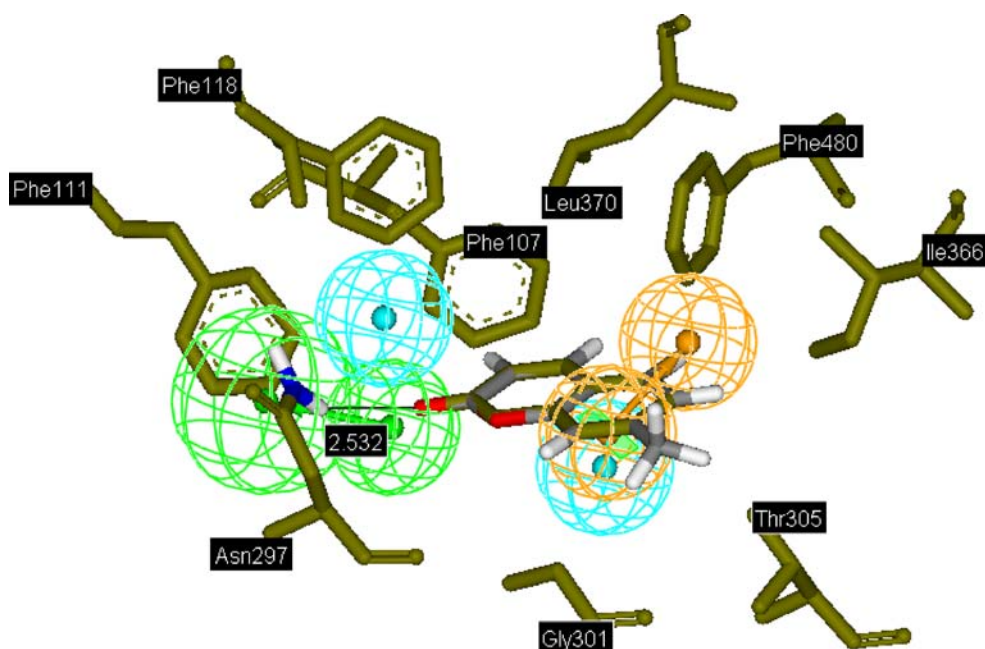


Fig. 7. The alignment of 7-methylcoumarin with coumarin in the CYP2A6-substrate co-complex structure (PDB code: 1Z10). The residues that constitute the active site and coumarin are denoted in green. The chemical features are depicted in Fig. 1.

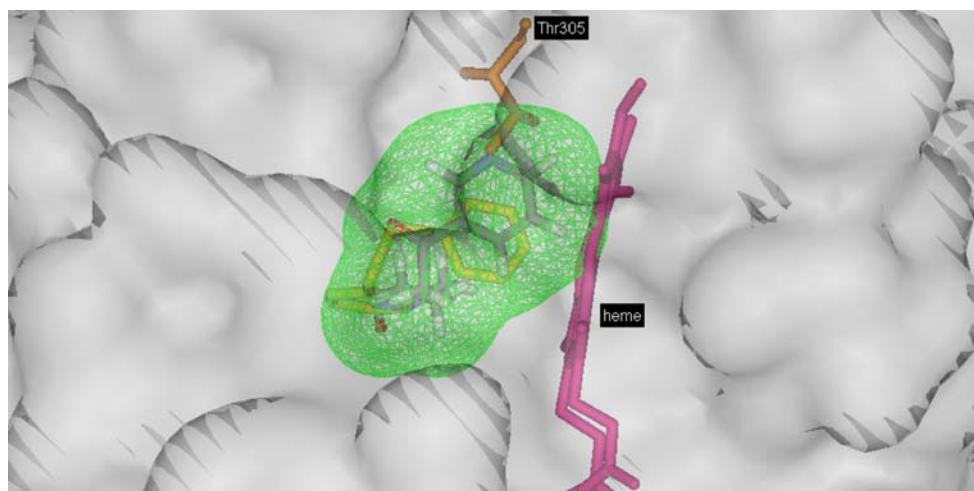


Fig. 8. The alignment of NNK with DPD in the CYP2A6-inhibitor co-complex structure (PDB code: 2FDY). The heme, Thr305 and DPD are represented in *magenta, brown and yellow*, respectively. The *green meshed lobe* shows the molecular volume of NNK.

of 7-methylcoumarin and that of Phe-480, suggesting that the π - π stacking is most likely due to the interaction between 7-methylcoumarin and Phe-480 instead of Phe-107. In addition, it can be observed from Fig. 7 that the ring aromatic chemical feature of Hypo C points to Phe-480 instead of Phe107.

In general, the geometry of the four-feature pharmacophore model Hypo C was found to be consistent with structural analyses of CYP2A6-coumarin crystal as mentioned above, suggesting that Hypo C is a plausible model to describe the interactions between 7-methylcoumarin and CYP2A6.

To further investigate how the CYP2A6 will adopt conformation change in order to accommodate more bulky molecules, NNK was placed in the active site of DPD-CYP2A6 co-complex since NNK is the most bulky and flexible molecule among 33 molecules included in this study and the DPD-CYP2A6 co-complex has the largest active site among all published structures. Unlike 7-methylcoumarin, which was compared with bound ligand by selecting the conformation matched by Hypo C as mentioned above, the mapped conformation of NNK by Hypo A was selected since the prediction by Hypo A yielded the lowest error (Table II). The variation in pharmacophore hypothesis selection by both molecules suggests that CYP2A6 enzyme can adopt different conformations to interact with structurally different substrates. Since NNK and DPD are structurally distinct, the selected NNK conformation was rigidly aligned against bound DPD using the *Cerius²* package (Accelrys, San Diego, CA). Fig. 8 displays the overlay of NNK and bound DPD in the enzyme-inhibitor co-complex. It can be observed that NNK is highly flexible that is in consistent with general postulate that not all of CYP2A6 substrates are planar (67). Most importantly, it can be observed that the molecular volume of aligned NNK collides with Thr305 and heme of CYP2A6 as shown in Fig. 8, rendering the fact that the active site of CYP2A6-DPD co-complex is too small for NNK despite the fact that the binding pocket of CYP2A6-DPD co-complex is the largest among all published structures. As a result, CYP2A6 protein has to change its conformation by

expanding its active site in order to accommodate more bulky NNK, giving rise to a new conformation with more spacious binding pocket, which is possible to be explored by time-consuming crystallization or computationally expensive QM MD calculations and the ensemble of protein conformation can be further extended accordingly so that the interactions between ligand and CYP2A6 can be more accurately modeled. The analog-based PhE/SVM scheme, on the other hand, can accurately predict the interactions between structurally distinct ligands and CYP2A6 by taking into account the protein plasticity using PhE without spending a lot of computational time.

CONCLUSION

An *in silico* model, based on the combination of pharmacophore ensemble, which takes into account protein plasticity while interacting with structurally distinct small molecules, and support vector machine, which provides robust and fast regression, has been built to accurately predict the interactions between CYP2A6 and those molecules in the training set, test set and external validation sets, with excellent predictability and statistical significance. As a result, it can be asserted, based on the facts mentioned above, that this PhE/SVM model can be employed as a tool for predictions and a device for high-throughput screening and data mining to facilitate drug discovery by reducing the attrition rates due to adverse side effects as well as by designing small molecule therapies for smoke cessation and chemoprevention of CYP2A6-associated cancers.

ACKNOWLEDGMENTS

This work was supported by the National Science Council, Taiwan. Parts of calculations were performed at the National Center for High-Performance Computing, Taiwan. The authors are grateful to Dr. G. H. Hakimelahi for reading the manuscript.

REFERENCES

1. D. F. Lewis, P. P. Tamburini, and G. G. Gibson. The interaction of a homologous series of hydrocarbons with hepatic cytochrome P-450. Molecular orbital-derived electronic and structural parameters influencing the haemoprotein spin state. *Chem. Biol. Interact.* **58**:289–299 (1986). doi:10.1016/S0009-2797(86)80104-1.
2. G. F. Roberts, I. Mehta, and M. Murray. Inhibition of oxidative drug metabolism by orphenadrine: *in vitro* and *in vivo* evidence for isozyme-specific complexation of cytochrome P-450 and inhibition kinetics. *Mol. Pharmacol.* **35**:736–743 (1989).
3. L. M. Forrester, C. J. Henderson, M. J. Glance, D. J. Back, B. K. Park, S. E. Ball, N. R. Kitteringham, A. W. McLaren, J. S. Miles, and P. Skett. Relative expression of cytochrome P450 isoenzymes in human liver and association with the metabolism of drugs and xenobiotics. *Biochem. J.* **281**:359–368 (1992).
4. M. VandenBranden, S. A. Wrighton, S. Ekins, J. S. Gillespie, S. N. Binkley, B. J. Ring, M. G. Gadberry, D. C. Mullins, S. C. Strom, and C. B. Jensen. Alterations of the catalytic activities of drug-metabolizing enzymes in cultures of human liver slices. *Drug Metab. Dispos.* **26**:1063–1068 (1998).
5. D. F. Lewis. Human cytochromes P450 associated with the phase I metabolism of drugs and other xenobiotics: a compilation of substrates and inhibitors of the CYP1, CYP2 and CYP3 families. *Curr. Med. Chem.* **10**:1955–1972 (2003). doi:10.2174/0929867033456855.
6. S. Micuda, L. Mundlova, E. Anzenbacherova, P. Anzenbacher, J. Chladek, L. Fuksa, and J. Martinkova. Inhibitory effects of memantine on human cytochrome P450 activities: prediction of *in vivo* drug interactions. *Eur. J. Clin. Pharmacol.* **60**:583–589 (2004). doi:10.1007/s00228-004-0825-1.
7. J. P. Villeneuve, and V. Pichette. Cytochrome P450 and liver diseases. *Curr. Drug Metab.* **5**:273–282 (2004). doi:10.2174/1389200043335531.
8. R. L. Walsky, and R. S. Obach. Validated assays for human cytochrome P450 activities. *Drug Metab. Dispos.* **32**:647–660 (2004). doi:10.1124/dmd.32.6.647.
9. T. Niwa, T. Shiraga, I. Ishii, A. Kagayama, and A. Takagi. Contribution of human hepatic cytochrome P450 isoforms to the metabolism of psychotropic drugs. *Biol. Pharm. Bull.* **28**:1711–1716 (2005). doi:10.1248/bpb.28.1711.
10. V. G. Divakaran, and A. T. Murugan. Polypharmacy: an undervalued component of complexity in the care of elderly patients. *Eur. J. Intern. Med.* **19**:225–226 (2008). doi:10.1016/j.ejim.2007.08.002.
11. D. W. Nebert, and D. W. Russell. Clinical importance of the cytochromes P450. *Lancet* **360**:1155–1162 (2002). doi:10.1016/S0140-6736(02)11203-7.
12. S. G. Bell, N. Hoskins, C. J. C. Whitehouse, and L. L. Wong. Design and engineering of cytochrome P450 systems. In A. Sigel, H. Sigel, and R. K. O. Sigel (eds.), *The Ubiquitous Roles of Cytochrome P450 Proteins*, Vol. 3, Wiley, Chichester, 2007, pp. 437–476.
13. T. Shimada, H. Yamazaki, and F. P. Guengerich. Ethnic-related differences in coumarin 7-hydroxylation activities catalyzed by cytochrome P450A6 in liver microsomes of Japanese and Caucasian populations. *Xenobiotica* **26**:395–403 (1996).
14. J. M. Tredger, and S. Stoll. Cytochromes p450—their impact on drug treatment. *Hosp. Pharm.* **9**:167–173 (2002).
15. C. Rodriguez-Antona, and M. Ingelman-Sundberg. Cytochrome P450 pharmacogenetics and cancer. *Oncogene* **25**:1679–1691 (2006). doi:10.1038/sj.onc.1209377.
16. Y. Kaida, N. Inui, T. Suda, H. Nakamura, H. Watanabe, and K. Chida. The CYP2A6*4 allele is determinant of S-1 pharmacokinetics in Japanese patients with non-small-cell lung cancer. *Clin. Pharmacol. Ther.* **83**:589–594 (2008). doi:10.1038/sj.clpt.6100484.
17. K. Ikeda, K. Yoshisue, E. Matsushima, S. Nagayama, K. Kobayashi, C. A. Tyson, K. Chiba, and Y. Kawaguchi. Bio-activation of Tegafur to 5-Fluorouracil is catalyzed by cytochrome P-450 2A6 in human liver microsomes *in vitro*. *Clin. Cancer Res.* **6**:4409–4415 (2000).
18. J. Hukkanen, P. Jacob III, and N. L. Benowitz. Metabolism and disposition kinetics of nicotine. *Pharmacol. Rev.* **57**:79–115 (2005). doi:10.1124/pr.57.1.3.
19. A. M. Lee, and R. F. Tyndale. Drugs and genotypes: how pharmacogenetic information could improve smoking cessation treatment. *J. Psychopharmacol.* **20**:7–14 (2006). doi:10.1177/1359786806066039.
20. M. K. Ho, and R. F. Tyndale. Overview of the pharmacogenomics of cigarette smoking. *Pharmacogenomics. J.* **7**:81–98 (2007). doi:10.1038/sj.tpj.6500436.
21. O. Pelkonen, A. Rautio, H. Raunio, and M. Pasanen. CYP2A6: a human coumarin 7-hydroxylase. *Toxicology* **144**:139–147 (2000). doi:10.1016/S0300-483X(99)00200-0.
22. J. A. G. Agúndez. Cytochrome P450 gene polymorphism and cancer. *Curr. Drug Metab.* **5**:211–224 (2004). doi:10.2174/1389200043335621.
23. D. W. Nebert, and T. P. Dalton. The role of cytochrome P450 enzymes in endogenous signalling pathways and environmental carcinogenesis. *Nat. Rev. Cancer* **6**:947–960 (2006). doi:10.1038/nrc2015.
24. A. Poso, J. Gynther, and R. Juvonen. A comparative molecular field analysis of cytochrome P450 2A5 and 2A6 inhibitors. *J. Comput.-Aided Mol. Des.* **15**:195–202 (2001). doi:10.1023/A:1008102217770.
25. A. Asikainen, J. Tarhanen, A. Poso, M. Pasanen, E. Alhava, and R. O. Juvonen. Predictive value of comparative molecular field analysis modelling of naphthalene inhibition of human CYP2A6 and mouse CYP2A5 enzymes. *Toxicol. Vitro* **17**:449–455 (2003). doi:10.1016/S0887-2333(03)00065-1.
26. M. Rahnasto, H. Raunio, A. Poso, C. Wittekindt, and R. O. Juvonen. Quantitative structure–activity relationship analysis of inhibitors of the nicotine metabolizing CYP2A6 enzyme. *J. Med. Chem.* **48**:440–449 (2005). doi:10.1021/jm049536b.
27. M. Rahnasto, C. Wittekindt, R. O. Juvonen, M. Turpeinen, A. Petsalo, O. Pelkonen, A. Poso, G. Stahl, H. D. Holtje, and H. Raunio. Identification of inhibitors of the nicotine metabolising CYP2A6 enzyme—an *in silico* approach. *Pharmacogenomics. J.* **8**:328–338 (2008). doi:10.1038/sj.tpj.6500481.
28. D. F. V. Lewis, M. Dickins, B. G. Lake, P. J. Eddershaw, M. H. Tarbit, and P. S. Goldfarb. Molecular modelling of the human cytochrome P450 isoform CYP2A6 and investigations of CYP2A substrate selectivity. *Toxicology* **133**:1–33 (1999). doi:10.1016/S0300-483X(98)00149-8.
29. D. F. V. Lewis. Homology modelling of human CYP2 family enzymes based on the CYP2C5 crystal structure. *Xenobiotica* **32**:305–323 (2002). doi:10.1080/00498250110112015.
30. D. F. V. Lewis, B. G. Lake, M. Dickins, and P. S. Goldfarb. Homology modelling of CYP2A6 based on the CYP2C5 crystallographic template: enzyme–substrate interactions and QSARs for binding affinity and inhibition. *Toxicol. in Vitro* **17**:179–190 (2003). doi:10.1016/S0887-2333(02)00132-7.
31. S. Sansen, M. H. Hsu, C. D. Stout, and E. F. Johnson. Structural insight into the altered substrate specificity of human cytochrome P450 2A6 mutants. *Arch. Biochem. Biophys.* **464**:197–206 (2007). doi:10.1016/j.abb.2007.04.028.
32. R. Arimoto. Computational models for predicting interactions with cytochrome P450 enzyme. *Curr. Top. Med. Chem.* **6**:1609–1618 (2006). doi:10.2174/156802606778108951.
33. A. M. Davis, and S. J. Teague. Hydrogen bonding, hydrophobic interactions, and failure of the rigid receptor hypothesis. *Angew. Chem. Int. Ed. Engl.* **38**:736–749 (1999). doi:10.1002/(SICI)1521-3773(19990315)38:6<736::AID-ANIE736>3.0.CO;2-R.
34. C. F. Wong, and J. A. McCammon. Protein flexibility and computer-aided drug design. *Annu. Rev. Pharmacol. Toxicol.* **43**:31–45 (2003). doi:10.1146/annurev.pharmtox.43.100901.140216.
35. M. G. McCammon, and C. V. Robinson. Structural change in response to ligand binding. *Curr. Opin. Chem. Biol.* **8**:60–65 (2004). doi:10.1016/j.cbpa.2003.11.005.
36. A. Verras, I. D. Kuntz, and P. R. O. de Montellano. Cytochrome P450 enzymes: Computational approaches to substrate prediction. In D. C. Spellmeyer (ed.), *Annual Reports in Computational Chemistry*, Vol. 2, Elsevier, Oxford, UK, 2006, pp. 171–195.
37. T. L. Poulos, and Y. T. Meharena. Structures of P450 proteins and their molecular phylogeny. In A. Sigel, H. Sigel, and R. K.

- O. Sigel (eds.), *The Ubiquitous Roles of Cytochrome P450 Proteins*, Vol. 3, Wiley, Chichester, 2007, pp. 57–96.
38. M. Ekroos, and T. Sjögren. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U. S. A.* **103**:13682–13687 (2006). doi:10.1073/pnas.0603236103.
 39. Y. Zhao, M. A. White, B. K. Muralidhara, L. Sun, J. R. Halpert, and C. D. Stout. Structure of microsomal cytochrome P450 2B4 complexed with the antifungal drug bifonazole: Insight into P450 conformational plasticity and membrane interaction. *J. Biol. Chem.* **281**:5973–5981 (2006). doi:10.1074/jbc.M511464200.
 40. M. R. Wester, E. F. Johnson, C. Marques-Soares, P. M. Dansette, D. Mansuy, and C. D. Stout. Structure of a substrate complex of mammalian cytochrome P450 2C5 at 2.3 Å resolution: Evidence for multiple substrate binding modes. *Biochemistry* **42**:6370–6379 (2003). doi:10.1021/bi0273922.
 41. V. M. Popov, W. A. Yee, and A. C. Anderson. Towards in silico lead optimization: scores from ensembles of protein/ligand conformations reliably correlate with biological activity. *Proteins* **66**:375–387 (2007). doi:10.1002/prot.21201.
 42. J. K. Yano, M.-H. Hsu, K. J. Griffin, C. D. Stout, and E. F. Johnson. Structures of human microsomal cytochrome P450 2A6 complexed with coumarin and methoxsalen. *Nat. Struct. Mol. Biol.* **12**:822–823 (2005). doi:10.1038/nsmb971.
 43. J. K. Yano, T. T. Denton, M. A. Cerny, X. Zhang, E. F. Johnson, and J. R. Cashman. Synthetic inhibitors of cytochrome P-450 2A6: inhibitory activity, difference spectra, mechanism of inhibition, and protein cocrystallization. *J. Med. Chem.* **49**:6987–7001 (2006). doi:10.1021/jm060519r.
 44. J. Dundas, Z. Ouyang, J. Tseng, A. Binkowski, Y. Turpaz, and J. Liang. CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **34**:W116–W118 (2006). doi:10.1093/nar/gkl282.
 45. I. B. Bersuker, M. K. Leong, J. E. Boggs, and R. S. Pearlman. A method of combined quantum mechanical (QM)/molecular mechanics (MM) treatment of large polyatomic systems with charge transfer between the QM and MM fragments. *Int. J. Quantum Chem.* **63**:1051–1063 (1997).
 46. R. A. Friesner, and V. Guallar. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* **56**:389–427 (2005). doi:10.1146/annurev.physchem.55.091602.094410.
 47. M. K. Leong. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. *Chem. Res. Toxicol.* **20**:217–226 (2007). doi:10.1021/tx060230c.
 48. B. Ma, M. Shatsky, H. J. Wolfson, and R. Nussinov. Multiple diverse ligands binding at a single protein site: a matter of pre-existing populations. *Protein Sci.* **11**:184–197 (2002). doi:10.1110/ps.21302.
 49. M. K. Leong, and T.-H. Chen. Prediction of cytochrome P450 2B6-substrate interactions using pharmacophore ensemble/support vector machine (PhE/SVM) approach. *Med. Chem.* **4**:396–406 (2008). doi:10.2174/157340608784872226.
 50. T. J. Smith, A. M. Liao, Y. Liu, A. B. Jones, L. M. Anderson, and C. S. Yang. Enzymes involved in the bioactivation of 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone in patas monkey lung and liver microsomes. *Carcinogenesis* **18**:1577–1584 (1997). doi:10.1093/carcin/18.8.1577.
 51. R. O. Juvonen, J. Gynther, M. Pasanen, E. Alhava, and A. Poso. Pronounced differences in inhibition potency of lactone and non-lactone compounds for mouse and human coumarin 7-hydroxylases (CYP2A5 and CYP2A6). *Xenobiotica* **30**:81–92 (2000). doi:10.1080/004982500237848.
 52. M. Rahnasto, H. Raunio, A. Poso, and R. O. Juvonen. More potent inhibition of human CYP2A6 than mouse CYP2A5 enzyme activities by derivatives of phenylethylamine and benzaldehyde. *Xenobiotica* **33**:529–539 (2003). doi:10.1080/0049825031000085979.
 53. T. T. Denton, X. Zhang, and J. R. Cashman. Nicotine-related alkaloids and metabolites as inhibitors of human cytochrome P-450 2A6. *Biochem. Pharmacol.* **67**:751–756 (2004). doi:10.1016/j.bcp.2003.10.022.
 54. J. M. MacDougall, K. Fandrick, X. Zhang, S. Serafin, and J. R. Cashman. Inhibition of human liver microsomal (S)-nicotine oxidation by (–)-menthol and analogues. *Chem. Res. Toxicol.* **16**:988–993 (2003). doi:10.1021/tx0340551.
 55. F. P. Guengerich. Drug metabolism as catalyzed by human cytochrome P450 systems. In A. Sigel, H. Sigel, and R. K. O. Sigel (eds.), *The Ubiquitous Roles of Cytochrome P450 Proteins*, Vol. 3, Wiley, Chichester, 2007, pp. 561–589.
 56. G. Chang, W. C. Guida, and W. C. Still. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **111**:4379–4386 (1989). doi:10.1021/ja00194a035.
 57. I. Kolossvary, and W. C. Guida. Low mode search. An efficient, automated computational method for conformational analysis: application to cyclic and acyclic alkanes and cyclic peptides. *J. Am. Chem. Soc.* **118**:5011–5019 (1996). doi:10.1021/ja952478m.
 58. W. C. Still, A. Tempczyk, R. C. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**:6127–6129 (1990). doi:10.1021/ja00172a038.
 59. T. A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **17**:490–519 (1996). doi:10.1002/(SICI)1096-987X(199604)17:5<490::AID-JCC1>3.0.CO;2-P.
 60. C.-C. Chang, and C.-J. Lin. LIBSVM: A Library for Support Vector Machines, version 2.81. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2005.
 61. V. Kecman. *Learning and soft computing: support vector machines, neural networks and fuzzy logic models*. MIT Press, Cambridge, MA, 2001.
 62. L. Breiman, and P. Spector. Submodel selection and evaluation in regression: the X-random case. *Int. Statist. Rev.* **60**:291–319 (1992). doi:10.2307/1403680.
 63. A. Golbraikh, and A. Tropsha. Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *J. Comput.-Aided Mol. Des.* **16**:357–369 (2002). doi:10.1023/A:1020869118689.
 64. E. Higashi, M. Nakajima, M. Katoh, S. Tokudome, and T. Yokoi. Inhibitory effects of neurotransmitters and steroids on human CYP2A6. *Drug Metab. Dispos.* **35**:508–514 (2007). doi:10.1124/dmd.106.014084.
 65. R. Gnanadesikan, and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**:81–124 (1972). doi:10.2307/2528963.
 66. A. C. Wallace, R. A. Laskowski, and J.M. Thornton. LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Prot. Eng.* **8**:127–134 (1995). doi:10.1093/protein/8.2.127.
 67. D. F. V. Lewis. On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics: towards the prediction of human P450 substrate specificity and metabolism. *Biochem. Pharmacol.* **60**:293–306 (2000). doi:10.1016/S0006-2952(00)00335-X.