Generalized Proteochemometric Model of Multiple Cytochrome P450 Enzymes and Their Inhibitors

Aleksejs Kontijevskis,[†,‡] Jan Komorowski,[‡] and Jarl E. S. Wikberg*[,†]

Department of Pharmaceutical Biosciences and Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden

Received March 16, 2008

Cytochrome P450 enzymes are a superfamily of heme-containing enzymes responsible for the oxidation of structurally diverse chemical compounds. Inhibition of CYP enzymes is probably the most common mechanism underlying acute drug toxicity, loss of therapeutic drug efficacy, and drug-drug interactions. The presence of polymorphic genetic variants of CYPs among the population makes it difficult to foresee undesired effects of drugs and is a common cause of drug candidate failure. Computational models that can predict early drug failures due to the inhibition of CYP isoforms can substantially reduce the cost of drug development. Although several computational models for CYP inhibition have been developed recently, all were constructed for one CYP isoform at a time, thus limiting their use for comprehensive analysis and generalizations to other CYP isoforms and polymorphisms. Here we report a novel approach based on the principles of proteochemometrics for the generalized concomitant modeling of multiple CYP isoforms and their inhibitors. We created a predictive and statistically valid proteochemometric model for CYP enzymes by combining data from a large number of publicly available reports that describe the interactions of 14 CYP enzyme subtypes and 375 structurally diverse inhibitors. Our results demonstrate that our model is capable of predicting the potential of new drug candidates to inhibit multiple CYP enzymes. Analysis of the CYP model also revealed molecular properties of CYP enzymes and xenobiotics that are important for CYP inhibition. This approach may aid in the selection of novel drug candidates that are unlikely to inhibit multiple CYP subtypes.

INTRODUCTION

Cytochrome P450 enzymes (CYPs) are a multigene family of heme-containing proteins that are involved in the oxidative metabolism of a large number of endogenous and exogenous compounds.[1,2] The human genome contains about 60 P450s, but more than 90% of all therapeutic drugs are metabolized by five main CYP isoforms: CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4.[3,4] Inhibition and/or induction of CYP isoforms by one drug often causes changes in the *in vivo* metabolism rates of other drugs. This can lead to various adverse drug interactions, e.g. loss in therapeutic efficacy or acute drug toxicity.[5,6]

Typically, properties of drug candidates that lead to costly failure relate to drug-drug interactions, absorption, metabolism, toxicity, and low efficacy and are only observed in clinical studies; thus drug failure may not be detected until late in the drug discovery and development process. The development of models that can predict drug failures due to the inhibition of CYP isoforms by xenobiotics at a preclinical stage can substantially minimize the cost of drug development by filtering away inadequate candidates early. However, the development of predictive models for CYPs is complicated by the fact that a number of polymorphic genetic variations in various CYP isoforms are observed among the population. In fact, variability of CYPs is the rule rather than the exception.[7] Moreover, for the most part, CYP mutations are silent until a new exogenous compound is encountered.[7] A single amino acid mutation in a CYP enzyme can have a profound effect on enzymatic activity and substrate specificity.[7,8] Thus, serious adverse drug reactions or a lack of drug activity may not be observed until a significant number of individuals with polymorphic CYP variations are tested.

A considerable number of models have been generated for CYP inhibitors, including models based on analyses of ligand binding, quantitative structure−activity relationships (QSAR), pharmacophore modeling, and 3D structure homology.[4,9−40] Some models were based on relatively small sets of similar ligands, and others were based on large proprietary data sets. However, each of these models considered only one particular CYP enzyme isoform at a time, and thus predictions could not be generalized to other CYP isoforms or polymorphisms.

Here we report a new strategy for the generalized concomitant analysis of multiple CYP enzymes and their inhibitors based on proteochemometrics. In proteochemometrics multiple protein and ligand parameters are correlated to protein−ligand interaction activities in a single general model that represents all studied entities.[41−48] We created and validated a highly predictive proteochemometric model for CYP enzymes by combining data from a large number of publicly available reports that describe the interactions of various CYP subtypes and structurally diverse inhibitors. Our results showed that the generalized CYP model could,

* Corresponding author phone: +46184714238; fax: +4618559718; e-mail: Jarl.Wikberg@farmbio.uu.se. Corresponding author address: Pharmaceutical Pharmacology, Box 591, BMC, SE-75124, Uppsala, Sweden.
† Department of Pharmaceutical Biosciences.
‡ Linnaeus Centre for Bioinformatics.

10.1021/ci8000953 CCC: $40.75
Published on Web 08/12/2008

with good accuracy, predict inhibition activities for new CYP enzyme isoforms and new compounds (i.e., predictions for data not included in the model occurred with RMSEPs 0.45−1.53 log units). Moreover, model analysis revealed some general enzyme and compound properties that determine their inhibitory potentials. The proposed strategy is generally applicable and may be scaled up for fast computational screenings of xenobiotics on multiple CYP subtypes and their polymorphic variants. This approach may aid in the selection of novel drug candidates that are unlikely to inhibit multiple CYP subtypes.

METHODS

Data for Cytochrome P450 Enzyme Inhibitors. The data were collected from various literature sources and comprised 798 observations (see Supporting Information Table S1). The data set consisted of the inhibitory IC_{50} activity of 375 structurally diverse drugs and "drug-like" molecules ("ligands") on the following CYP enzymes: CYP1A1, CYP1A2, CYP1B1, CYP2A5, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4, CYP3A5, and CYP3A7. All CYPs were human, except CYP2A5 which was of murine origin. They were all included in order to expand the "chemical" space of the CYPs as much as possible by use of available public data.

Description of Ligands. Ligands were drawn in ISIS/Draw (MDL Information Systems, Inc. San Leandro, CA), and a single conformation of their 3D coordinates was generated for each structure by the CORINA program (Molecular Networks, v. 3.4).[49] The compounds showed large chemical diversity, ranging from rigid to very flexible structures and spanning molecular weights from 80 to 850 daltons. Alignment independent GRIND descriptors (GRid-INdependent Descriptors) were calculated automatically with Almond software (Molecular Discovery, v. 3.3).[50,51] GRIND descriptors represent energy regions where amino acid groups of a protein could interact favorably with the ligand. The generation of GRIND descriptors involves three main steps: i) calculation of a set of molecular interaction fields (MIFs), ii) filtering MIFs, and iii) final transformation of the extracted MIF regions into GRIND variables. In this study we generated MIFs for DRY, N1, O, and TIP probes,[50,51] defined as follows: the DRY probe represents hydrophobic interactions, N1 (amide) and O (carbonyl) probes represent hydrogen bond donor and acceptor groups, respectively, and the TIP probe represents a "shape-field". All molecular interaction fields were computed using a grid spacing of 0.5 Å, with the grid extending 5 Å beyond the molecule. The maximum number of extracted nodes was set to 100, and the smoothing window was set to the default value of 0.8 grid units. Ten correlograms containing 82 variables each were calculated to explain the DRY-DRY, O−O, N1−N1, and TIP-TIP interactions (autocorrelograms) and the DRY-O, DRY-N1, O−N1, DRY-TIP, N1-TIP, and O-TIP interactions (cross-correlograms); this analysis resulted in 820 descriptors for each ligand. We removed 155 GRIND descriptors that contained only zero values for all ligands as they did not show any variation. This resulted in 665 remaining descriptors.

Description of the Cytochrome P450 Enzymes. The 14 cytochrome P450 enzyme sequences used in this study were aligned using the alignment available at the public database of cytochrome P450 enzymes (http://drnelson.utmem.edu/humP450.aln.html) (see also Supporting Information Figure S1). Among all 14 enzymes, 466 amino acids, representing 90−95% of the CYP sequences, could be fully aligned and were considered further. Of these, we selected only the nonconserved positions in the data set (438 amino acids); the importance of the conserved amino acids for the inhibitory activity could not be evaluated because there were insufficient examples of proteins with mutations in conserved positions in the CYP library. The sequence analysis revealed that a number of aligned amino acid positions had the same variance throughout the alignment. These "covarying" amino acid positions could be grouped into six different groups of two or more positions in each group, where the variability among the sequences was of a binary nature (i.e., groups A-F, see CYPs alignment in Supporting Information Figure S1). For example, group A consisted of 16 aligned positions as indicated in CYPs alignment (see Supporting Information Figure S1 for details); at each position one of two amino acids was observed for each CYP, i.e. the same amino acid was present in CYP3A4, CYP3A5, and CYP3A7 enzymes, while another was present among all other CYP isoforms. Because the 16 amino acids of group A were covariants, this group of amino acids was described by one binary descriptor; the value was set to "+1" if the amino acids were present in the CYP3A4 sequence and to "−1" otherwise. Describing other groups of covarying amino acids similarly resulted in a total of 6 binary descriptors corresponding to groups A-F. Further analysis of the alignment showed that there were 20 additional amino acid positions that varied between two amino acids, without any covariation with other amino acid positions. Each of these 20 amino acid positions was also described by a binary descriptor in the same way as described above (i.e., resulting in 20 descriptors in total). The remaining 388 amino acid positions were described by five z-scale[52] properties, yielding 1940 descriptors (388×5). These z-scales represented principal components for 26 physical and chemical properties measured in 87 natural and unnatural amino acids and were independent (i.e., being orthogonal) of each other. The z_1-scale roughly corresponds to amino acid hydrophobicity, z_2 to steric properties, and z_3-z_5 to polarizability, polarity, and electronic effects of amino acids, respectively. Thus, each CYP enzyme was described by $1940 + 6 + 20 = 1966$ total descriptors.

Description of Inhibition Experiments. In some publications the IC_{50} value for CYP enzyme inhibitions was given as "less" or "more" than a particular value. In those cases we used the actual value shown. Inhibition of CYP activities is substrate dependent, and therefore the IC_{50} value for the same CYP-inhibitor pair may vary with the use of different substrates.[19] However, the IC_{50} differences for such cases were generally small in the data set and rarely exceeded 0.5 decimal log units IC_{50}. For the cases in the data used herein where IC_{50} values had been determined using several different substrates for the same CYP-inhibitor pair, we used the IC_{50} value that was determined with the most commonly used substrate for a particular CYP subtype. Finally, all IC_{50} values for CYP-ligand interactions were converted to molar units and then transformed to decimal logarithms "$-\log(IC_{50})$" or pIC_{50}. The pIC_{50} values used herein spanned over eight log units (see Supporting Information Table S1 for details).

Preprocessing of Descriptors. The large numbers of CYP and ligand descriptors (1966 and 665, respectively) posed the danger of obtaining chance correlations in the subsequent modeling, especially once interaction terms were introduced. For example, the number of CYP-ligand cross-term descriptors would be the following: $1966 \times 820 = 1{,}612{,}120$, representing a large number of theoretical interactions compared to the number of actual observations in the data set (798). Moreover, many of the descriptors were covariant. In order to overcome these problems we reduced the number of ordinary descriptors by applying principal component analysis (PCA). PCA approximates the multivariate data by projecting it onto a lower dimensional variable space, called latent variables or principal components (PC). Components are computed iteratively and are orthogonal to each other. Prior to the PCA calculations, the descriptors of CYP enzymes and ligands were mean centered and scaled to unit variance. PCA was applied to the CYP enzyme descriptor block and ligand GRIND descriptor block separately. The resulting 13 CYP PCs (here termed CYP_{PC1}-CYP_{PC13}) were sufficient to explain 100% of the variance in the initial CYP descriptor block data set. The 40 PCs obtained for the ligand GRIND descriptors (Lig_{PC1}-Lig_{PC40}) explained 91% of the variance in the initial ligand descriptor block data set. These PCs were then used as CYP enzyme and ligand descriptors, respectively, in the subsequent modeling. PCA was performed using the SIMCA-P+ 11.5 program (http://www.umetrics.com).

Description of CYP-Ligand Interactions. Each CYP-ligand combination was described by 13 CYP descriptors and 40 ligand descriptors. In proteochemometrics the interactions between proteins and ligands may be described by protein−ligand cross-terms, here calculated by multiplying each one of the CYP descriptors with each one of the ligand descriptors. In addition to CYP × ligand cross-terms we also used CYP × CYP cross-terms (calculated as the products of two CYP descriptors) and ligand × ligand cross-terms (calculated as the products of two ligand descriptors). Prior to calculating cross-terms all descriptors had been mean centered and scaled to unit variance.

Multivariate Modeling. The preprocessed CYP and ligand descriptors and their interaction products (i.e., CYP × CYP, CYP × ligand and ligand × ligand cross-terms) were correlated in various combinations to the pIC_{50} values using partial least-squares projections to latent structures (PLS) modeling[53] until the best models were achieved according to state-of-the-art validation (see below). All descriptors were mean-centered and scaled to unit variance before generation of the PLS models. In PLS both the descriptor and response spaces are simultaneously projected to a lower dimensional hyperspace and constrained to find the maximal covariance between the components.[53] PLS was performed using SIMCA-P+ 11.5 software (http://www.umetrics.com).

To improve the model's performance further we used a variable selection procedure. In the proteochemometric model, the relative overall importance of the individual X variables with respect to a response Y variable was characterized by the variable's influence on projection, VIP: the sum of the variable's influences over all model dimensions.[54] The sum of squares of all the VIPs is equal to the number of terms in the model, and the average VIP is equal to 1. VIPs can be used to compare the relative importance of terms and cross-terms in order to find those that best explain the response variable Y. Exclusion of terms with small VIPs ($VIP < 0.6$) can improve the validity and performance of PLS models.[54]

Validation of Models. The goodness of fit for a PLS model was assessed by R^2, which is the fraction of total variation of the response that could be explained by the model. We also computed the root-mean-square error of the model estimate (RMSEE) in order to measure the internal error within the model.

All models generated were validated by 7-fold cross-validation (CV) and response permutation validation. In CV the data were randomly divided into 7 equal parts. Each part was then iteratively excluded from the data set, and pIC_{50} values were predicted based on the model constructed on the remaining 6 parts of the data. The goodness of CV was assessed by the Q^2 measure (Simca-P+ 11.0 User Guide and Tutorial, 2005, Umetrics AB). The CV procedure was repeated 20 times for each model on different data subdivisions in order to obtain an estimate of the Q^2 variability.

In response permutation validation, the pIC_{50} values of the data set were randomly permuted (100 sets of permuted pIC_{50}s were used herein). New models were then built on each permuted data set, and their R^2 and Q^2 were estimated. The correlation coefficients between the original R^2 and Q^2 and the 100 permuted R^2 and Q^2 values were computed. In addition, two corresponding linear regression lines were calculated (i.e., one for R^{2s} and one for the Q^2s), and their intercepts (iR^2 and iQ^2) with the zero correlation coefficient line were estimated. These intercept values indicated the R^2 and Q^2 of random response data. The criteria for a statistically valid model based on biological data were as follows: $R^2 > 0.7$, $Q^2 > 0.4$, $iR^2 < 0.4$, and $iQ^2 < 0.05$ (Simca-P+ 11.0 User Guide and Tutorial, 2005, Umetrics AB).[54,55]

We also used two additional validation tests to assess the generalizability of the CYP_M6 model. In the first test the data set was randomly split into two parts (90% and 10%). The model was built on 90% of the data and was then used to predict pIC_{50} for the excluded 10% of the data, and the root-mean-square error of prediction (RMSEP) was computed. This test was repeated 100 times, and a mean value of all RMSEPs and its standard error were calculated. The second "leave-one-CYP-isoform-out" validation was more rigorous. For this test we excluded inhibitory data for each CYP enzyme subtype, one at a time, and constructed PLS models on the data for the remaining 13 CYP enzymes. We then predicted PCA scores for the excluded CYP enzyme (based on the PCA model for the CYPs included in the training set) and PCA scores for the excluded ligands (based on the PCA model for the ligands included in the training set). The PLS model constructed on the data for 13 CYP enzymes was used to predict pIC_{50} values for the excluded CYP isoform. We also computed RMSEP, the correlation coefficient r to capture the similarity between observed and predicted values, and its statistical significance p. The p-value is the probability that a given correlation or greater (in the positive direction only) would be observed if there were no real linear relationship between the observed and predicted pIC_{50} values. We chose to use a one-sided significance test. The correlation and all significance tests were performed with a proprietary add-in to the Excel program (Microsoft).

Interpretation of Models. The regression equation for the CYP_M6 model is expressed as follows

$$pIC_{50} = \overline{pIC_{50}} + \sum_{i=1}^{I} k_i CYP_{PC_i} + \sum_{j=1}^{J} k_j Lig_{PC_j} + \sum_{n=1}^{N} CYP_{PC_n}\left(\sum_{m=1}^{M} k_{nm} Lig_{PCm}\right) + \sum_{i=1}^{I} CYP_{PC_i}\left(\sum_{a=2,a>i}^{A} k_{ai} CYP_{PC_a}\right) + \sum_{j=1}^{J} Lig_{PC_j}\left(\sum_{b=2,b>j}^{B} k_{bj} Lig_{PC_b}\right) \quad (1)$$

where $\overline{pIC_{50}}$ indicates the average pIC_{50} of the data set; I, J, $N \times M$, $I \times A$, and $J \times B$ correspond to the number of CYP, ligand descriptors, CYP × ligand, CYP × CYP, and ligand × ligand cross-terms respectively; and k_i, k_j, k_{nm}, k_{ai}, and k_{bj} correspond to PLS coefficients for CYP, ligand, CYP × ligand, CYP × CYP, and ligand × ligand descriptor blocks, respectively.

The importance of each descriptor for inhibition of CYP enzymes was assessed by a PLS coefficient plot for the CYP_M6 model. The PLS coefficient plot indicated the size and direction of the impact a descriptor had on the predicted pIC_{50} values. A large coefficient for an ordinary ligand descriptor or a ligand × ligand cross-term in eq 1 indicated that the presence or absence of one particular or several molecular interaction fields (MIFs) in one ligand would have a large influence on the inhibition constant of a CYP, regardless of the isoform involved. (This is because neither ligand descriptors nor ligand × ligand cross-terms depend on CYP descriptors). On the other hand, a large PLS coefficient for a ligand × CYP cross-term indicated a favorable interaction between a particular ligand MIF and a particular set of CYP enzyme amino acids, i.e. the ability of a particular ligand to interact with and inhibit a particular CYP enzyme. This allows comparisons in ligand potencies. For example, the difference between ligand *U*'s inhibition of some CYP enzyme *B* and some CYP enzyme *C* could be expressed as follows:

$$\Delta pIC_{50}(Lig_U CYP_{BC}) = \sum_{i=1}^{I} k_i\left(CYP_{BPC_i} - CYP_{CPC_i}\right) + \sum_{n=1}^{N}\left(CYP_{BPC_n} - CYP_{CPC_n}\right)\left(\sum_{m=1}^{M} k_{nm} Lig_{UPC_m}\right) + \sum_{i=1}^{I}\left(CYP_{BPC_i} - CYP_{CPC_i}\right)\sum_{a=2,a>i}^{A} k_{ai}\left(CYP_{BPC_a} - CYP_{CPC_a}\right) \quad (2)$$

Eq 2 shows that the difference in ligand potency is due to differences in the CYP enzymes' descriptors as well as the size of the corresponding regression coefficients. If the regression coefficients k for CYP ordinary descriptors or CYP × CYP cross-term descriptors were small, then the ΔpIC_{50} would be also small, and ligand U would have similar inhibition potency for both CYP enzymes.

We used projected scores (principal components) as descriptors of the CYP enzymes and ligands in the PLS model rather than the original descriptors. To find out which original descriptors contributed most to the principal components we analyzed the initial CYP enzyme and ligand PCA models.

RESULTS

The aim of the study was to build a general predictive proteochemometrics model for a large number of CYP enzymes and structurally diverse inhibitors. CYP models were generated by using partial least-squares regression to determine how various combinations of the molecular descriptors of CYP enzymes, their inhibitors, and cross-terms correlated to the inhibition constant pIC_{50} (Table 1). Cross-validation assessments showed that the introduction of cross-terms into the multivariate modeling significantly improved model performance (see models CYP_M2 to CYP_M5 in Table 1). However, the inclusion of all the cross-term descriptors simultaneously into the CYP_M5 model resulted in a relatively high iR^2 parameter, indicating potential overfitting. We therefore performed variable selection by removing all the CYP_M5 model descriptors that had a VIP < 0.6 (see the Methods section for details). The resulting CYP_M6 model outperformed all the other models and was chosen for further validation (Figure 1).

The CYP_M6 model was validated by demonstrating its ability to predict new data using a random 90% of the data to build the model and then predicting the remaining 10% of the data. This test showed that the model was robust and could predict new data with a RMSEP of 0.73 ± 0.08 log units. In a second "leave-one-CYP-isoform-out" validation, all the data for one particular CYP enzyme were excluded, and the model was constructed on the remaining data. This test yielded statistically significant and accurate predictions of pIC_{50} values for the majority of the excluded CYP enzyme isoforms (RMSEPs 0.45−1.53 log units; see Figure 2). Predictions of some excluded CYP-ligand pairs were less accurate than for the majority of observations used in the tests (e.g., see some observations for CYP2D6 and CYP1A2 in Figure 2). This is expected and can be explained by a decreased scope of the generalized CYP model when a large portion of the data is excluded. Although the inhibition data for CYP1B1, CYP2B6 and CYP2E1 enzymes were predicted with small RMSEP, the statistical significance (*p-value*) of the correlation r between predicted and observed pIC_{50} values exceeded the $p < 0.05$ limit (data not shown). This can be attributed to the small number of observations in the test set for these CYPs. Overall, validation results demonstrate that we have constructed a generally valid model of multiple CYP enzymes that can be used for predicting the potencies of new drug candidates for the inhibition of multiple CYP enzymes.

We used principal component descriptors rather than original CYP and ligand descriptors in the modeling. Although original descriptors would have been simpler to interpret, the number of cross-terms would substantially outnumber the available observations, make modeling computationally expensive, and increase the likelihood of overfitting. The principal component descriptors are independent of each other and compress the original descriptor space while essentially retaining information content.[45] However, as loadings of original variables indicate the extent to which these variables are influential in forming a PCA-based variable one can assign meanings to a principal component, e.g. in the way as is shown in Table 2.

To interpret the CYP_M6 model we used the PLS coefficient plot shown in Figure 3. The most important ligand principal component descriptors were Lig_{PC4}, Lig_{PC6}, Lig_{PC7},

Table 1. Validation Results of Generalized Cytochrome P450 Models[a]

models	R^2	$Q^2_{mean} \pm SE$	iR^2	iQ^2	RMSEE, pIC_{50}	NC	Descriptors used	no. of descriptors
CYP_M1	0.43	0.36 ± 0.01	0.05	−0.14	0.91	2	CYP, ligand	53
CYP_M2	0.45	0.39 ± 0.01	0.05	−0.19	0.89	4	CYP, ligand, CYP × CYP	131
CYP_M3	0.62	0.40 ± 0.01	0.26	−0.20	0.74	2	CYP, ligand, CYP × ligand	573
CYP_M4	0.59	0.44 ± 0.01	0.17	−0.22	0.77	6	CYP, ligand, ligand × ligand	833
CYP_M5	0.78	0.61 ± 0.01	0.42	−0.30	0.56	7	CYP, ligand, CYP × CYP, CYP × ligand, ligand × ligand	1431
CYP_M6	0.78	0.66 ± 0.01	0.29	−0.32	0.54	7	CYP, ligand, CYP × CYP, CYP × ligand, ligand × ligand	983

[a] CYP models (CYP_M1 - CYP_M6) were constructed by inclusion of different groups of descriptors (CYP, ligands and their cross-term descriptors) in various combinations. CYP × CYP, CYP × ligand, and ligand × ligand descriptors represent the cross-terms formed from the respective ordinary descriptors. All descriptors shown were used for the construction of the CYP_M1 - CYP_M5 models; however, only descriptors from the CYP_M5 model with VIP > 0.6 were used for the construction of the model CYP_M6. RMSEE denotes root-mean-square error of estimation. SE denotes the standard error of the Q^2_{mean}. NC indicates the number of significant PLS components used in the model construction.

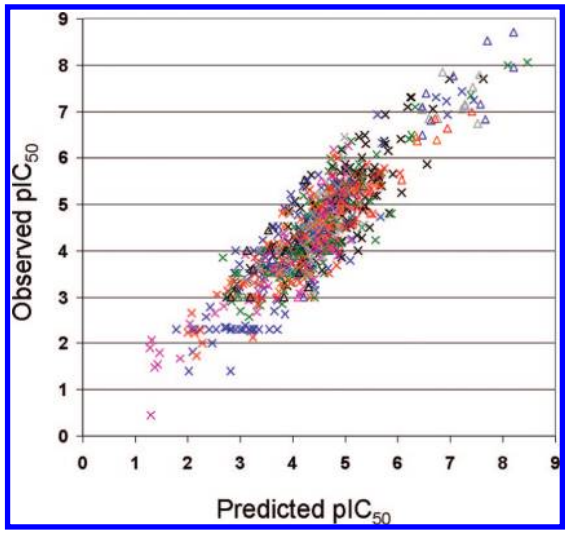

Figure 1. Goodness of fit for the CYP_M6 model. Observed versus predicted inhibition constants for the 14 CYP enzymes and 375 ligands (in total 798 CYP-ligand data observations), where predictions related to model-building data (R^2=0.78, Q^2=0.66). Colored crosses indicate observations of the CYP2D6 (black), CYP1A2 (blue), CYP3A4 (green), CYP2A6 (magenta), CYP2A5 (red), and CYP2C9 (orange) enzymes. Colored triangles correspond to observations of the CYP2C19 (gray), CYP2E1 (black), CYP1A1 (blue), CYP2C8 (green), CYP2B6 (magenta), CYP1B1 (gray), CYP3A5 (red), and CYP3A7 (orange) enzymes.

and Lig_{PC9}, while the most significant ligand × ligand cross-terms descriptors were $Lig_{PC1} \times Lig_{PC6}$, $Lig_{PC2} \times Lig_{PC6}$, $Lig_{PC6} \times Lig_{PC35}$, $Lig_{PC7} \times Lig_{PC35}$, and $Lig_{PC11} \times Lig_{PC33}$. Table 2 lists the original ligand GRIND descriptors that contributed most to the principal components. For interpretation, positive coefficient and descriptor products increase the inhibition potential, while negative products reduce the inhibition potential. For example, the coefficient for the Lig_{PC6} descriptor is large and negative (Figure 3). Thus a positive Lig_{PC6} descriptor (positive loading in Table 2) would yield a negative coefficient × descriptor product and would tend to decrease the compound's pIC_{50} values *over all* CYP enzymes. However, this product would be overcome if the cumulative effect of all the coefficient × cross-term products where Lig_{PC6} is involved yielded a larger positive number (see eq 1 in the Methods section). This is important information that could be used for lead compound optimization in drug design and help to clarify the factors that increase and reduce the likelihood of lead compounds' inhibition of CYPs.

The negative loadings in Table 2 indicate that Lig_{PC6} interactions with O nodes situated 15.6−16.4 Å from each other, with DRY nodes 4.0−8.0 Å apart, and with O−N1 nodes 16.0−16.8 Å apart would cause an increase in the inhibition activity and thus should be avoided for optimizing a drug candidate. On the other hand, compounds with N1-TIP nodes at a distance of 7.6−9.6 Å are positively loaded and would be more favorable. Similarly, according to the positive coefficient of the ligand descriptor Lig_{PC9} (Figure 3), shape-field TIP probes at distances 2.0−4.0 Å and 14.8−15.6 Å should be discouraged, while TIP probes 7.2−9.6 Å apart would reduce CYP inhibition and thus be favorable (Table 2). The practical use of these results can be exemplified with propranolol, which contains a naphthalene group that generates strong hydrophobic interactions with the DRY probe situated at 4.0−8.0 Å apart (propranolol has large original DRY-DRY descriptors at 4.0−8.0 Å, averaging 0.36 according to the ligand PCA model) (see Figure 4A). These DRY-DRY descriptors are responsible for some of the inhibitory activity of CYPs according to the model (i.e., pIC_{50} for CYP3A4 and CYP2D6 are 3.85 and 5.18, respectively, see Supporting Information Table S1). In order to reduce the inhibitory activity of propranolol one may substitute one benzene ring with a less hydrophobic group. For instance, the exchange of one aromatic ring by a metoxyethyl group leads to metoprolol, which has smaller DRY-DRY descriptors at all distances between 4.0 and 8.0 Å (averaging 0.29 according to the PCA model of the ligands) (see Figure 4B). Metoprolol is therefore a weaker CYP inhibitor than propranolol (pIC_{50} of metoprolol is 3.52 and 4.52 for CYP3A4 and CYP2D6, respectively) (see also Supporting Information Table S1).

Interpretation of the ligand × ligand cross-term descriptors is a more complex issue (see interpretation of models in the Methods section for details). Two principal component descriptors are involved in each ligand × ligand cross-term (this group of cross-terms has no dependence on CYP descriptors), thus the simultaneous presence or absence of several molecular interaction fields (MIF) impacts the ligand potencies over all CYP enzymes. For instance, since the coefficient for the $Lig_{PC1} \times Lig_{PC6}$ cross-descriptor is large

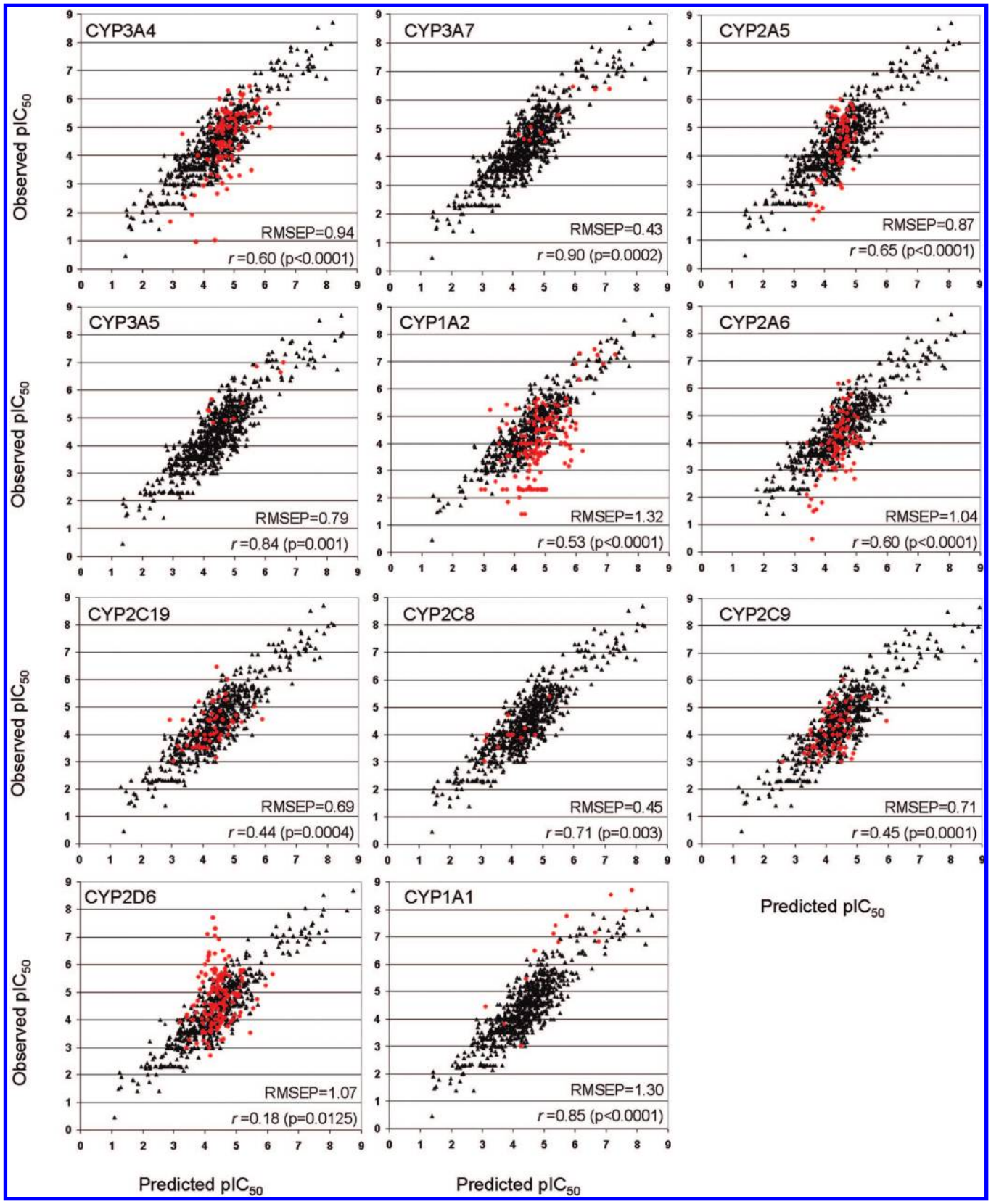

Figure 2. Leave-one-CYP-isoform-out validation of the CYP_M6 model. Each panel shows predictions of the CYP_M6 model constructed on the data set available in Supporting Information Table S1, but excluding data for one CYP enzyme isoform, as specified on each panel. Red bullets indicate the inhibition constant, pIC_{50}, predicted for the excluded CYP enzyme isoform. Black triangles correspond to the observed versus calculated model-building data. RMSEP is the root-mean-square error of the prediction; r is the correlation coefficient for the observations marked by red bullets, and the p-value for r is shown.

and positive, the descriptors Lig_{PC1} and Lig_{PC6} should be differently loaded in order to achieve the desired decrease in pIC_{50}. According to Table 2 this situation occurs when any of the MIFs, DRY-N1 (11.6−18.0 Å apart), DRY-O (14.0−14.4 Å apart), or N1-TIP (17.2−19.6 Å apart), are present in the compound (i.e., contributing positively to

Table 2. Ordinary CYP and Ligand Descriptors That Contribute Most to the Principal Components Revealed by the PLS Coefficient Plot of the CYP_M6 Model[a]

principal component descriptor	negative loadings	positive loadings
CYP_{pc1}	F60(z_1), R106(z_3), L133(z_1,z_3), T138(z_2), I193(z_3), V312(z_1,z_2), L365(z_2), L372(z_1)	C64(z_1), G73(z_4), D86(z_5), M145(z_4), Y179(z_2), R267(z_1), F315(z_3), Y318(z_2), E319(z_1), K377(z_2)
CYP_{pc2}	P43(z_2,z_3,z_4), Y75(z_3), R161(z_4), S188(z_5), P433(z_1,z_3,z_4), F462(z_2), G480(z_1) and covarying amino acids (I334, V358)	T92(z_5), T136(z_3), E307(z_5), L363(z_3), P43(z_1), G480(z_2), L481(z_1), P487(z_1)
CYP_{pc8}	L24(z_3), D174(z_3), N236(z_2), L248(z_5), L271(z_2,z_3,z_4), Q272(z_4), W391(z_3), I426(z_5), Y431(z_4), E485(z_2), V492(z_4)	N197(z_2), M1(z_2), L248(z_3,z_4), H323(z_3), N461(z_2), E485(z_3)
Lig_{pc1}	-	DRY-N1 (11.6−18 Å) DRY-O (14.0−14.4 Å) N1-TIP (17.2−19.6 Å)
Lig_{pc2}	-	DRY-DRY (23.2−26.0 Å) N1−N1 (26−26.4 Å) TIP-TIP (26.8−30.4 Å)
Lig_{pc4}	O-TIP (2.0−5.2 Å)	DRY-DRY (8.0−12.8 Å) DRY-TIP (15.6 Å)
Lig_{pc5}	-	O−N1 (21.6−26.4 Å) O−O (22.0−26.0 Å) O-TIP (24.4 Å)
Lig_{pc6}	O−O (15.6−16.4 Å) O−N1 (16−16.8 Å) DRY-DRY (4.0−8.0 Å)	N1-TIP (7.6−9.6 Å)
Lig_{pc7}	O−O (20.8−21.6 Å) DRY-O (23.2 Å)	O−N1 (24.0−26.4 Å) O−O (23.2−26.0 Å) O-TIP (25.2 Å)
Lig_{pc8}	DRY-DRY (2.0−2.8 Å) DRY-O (3.2−6.8 Å) DRY-TIP (5.6−6.8 Å)	N1-TIP (2.0−4.0 Å; 15.2 Å)
Lig_{pc9}	DRY-TIP (4.4−5.6 Å) TIP-TIP (7.2−9.6 Å)	TIP-TIP (2.0−4.0 Å; 14.8−15.6 Å)
Lig_{pc11}	O-TIP (2.8 Å) N1-TIP (4.0 Å)	DRY-O (2.0−2.4 Å) N1−N1 (4.8 Å) O−O (2.0−7.6 Å)
Lig_{pc33}	N1−N1 (15.6 Å) TIP-TIP (5.6−6.8 Å; 12.4−14.0 Å)	DRY-DRY (5.6 Å; 22.0−22.8 Å) DRY-TIP (4.4 Å) O−O (14.4−16 Å)
Lig_{pc35}	DRY-O (2.4−3.6 Å) DRY-TIP (3.6 Å) O−N1 (19.2−21.2 Å) O-TIP (15.2 Å)	DRY-DRY (8.8−9.6 Å) DRY-TIP (2.4 Å; 11.2 Å) O−O (18.4 Å) O-TIP (18.8−19.2 Å)

[a] Negative loadings indicate descriptors with the largest negative weight for the corresponding principal component. Positive loadings indicate descriptors with the largest positive weight for the respective principal component. The 20 descriptors that contributed most to each principal component are also shown. Amino acid designations correspond to the numbering of amino acids in the CYP3A4 sequence.

Lig_{PC1} descriptor) at the same time as any of the MIFs, O−O (15.6−16.4 Å apart), O−N1 (16.0−16.8 Å apart), or DRY-DRY (4.0−8.0 Å apart), are present in the same compound (contributing negatively to Lig_{PC6}).

A number of important CYP × ligand cross-terms are revealed by the PLS coefficients plot of the CYP_M6 model (Figure 3). In contrast to ligand × ligand cross-terms, where the presence or absence of particular MIFs in the ligand confer a potential decrease or increase in ligand activity over all CYP enzymes, CYP × ligand cross-descriptors reflect a ligand's selectivity for a particular CYP enzyme (see eq 2 in the Methods section). For example, the largest cross-term coefficient is the negative coefficient for the CYP_{PC2} × Lig_{PC6} cross-term, which suggests that if one of the cross-term descriptors is also negative, a ligand-CYP enzyme combination will form favorable interactions, leading to an increase in pIC_{50}. In other words the CYP enzymes with a large positive CYP_{PC2} descriptor (e.g., CYP1A1, CYP1A2, CYP1B1, or CYP2D6) would bind most favorably with ligands that contain MIFs O−O (15.6−16.4 Å apart), O−N1 (16−16.8 Å apart), or DRY-DRY (4.0−8.0 Å apart) which afford higher inhibition activities, whereas ligands with N1-TIP MIFs (7.6−9.6 Å apart) would bind less favorably (Table 2).

DISCUSSION

Here we have developed a general valid proteochemometrics model for predicting the interactions of CYP enzymes and their inhibitors. Although the experimental data utilized to populate the model came from different research groups, the z-scale and GRIND-descriptors make the model robust, and it appears to be highly useful for predicting IC_{50} values for new CYP enzymes and a variety of compounds. Our work, therefore, represents an important step toward the creation of general structure−activity relationship models for drug candidates and human CYP P450s involved in drug metabolism and may be useful in rational drug design.

We compared our results to earlier published computational models for inhibition of individual CYP enzymes (see

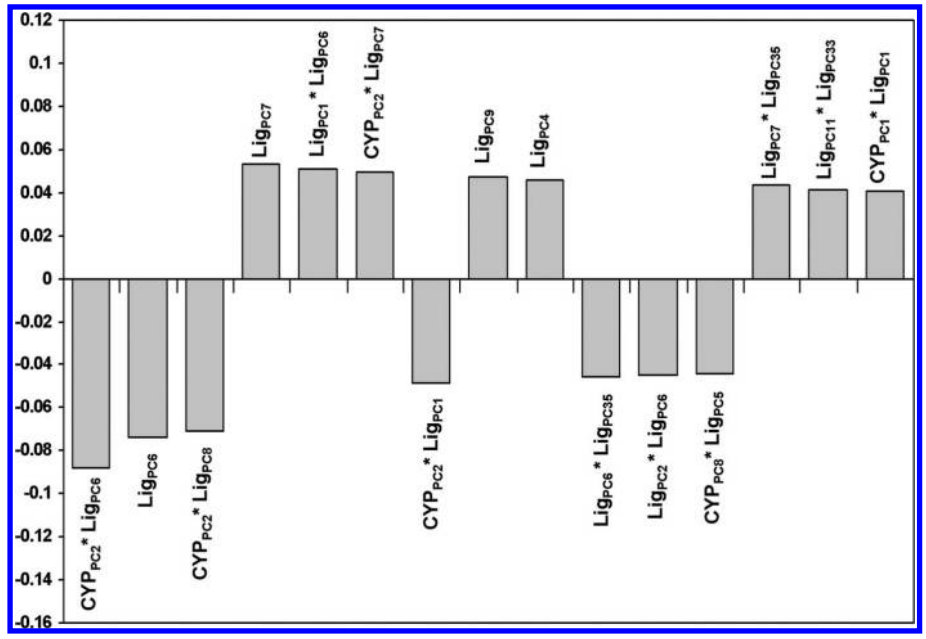

Figure 3. PLS coefficient plot of the CYP_M6 model. The PLS coefficients for the 15 most important descriptors are shown.

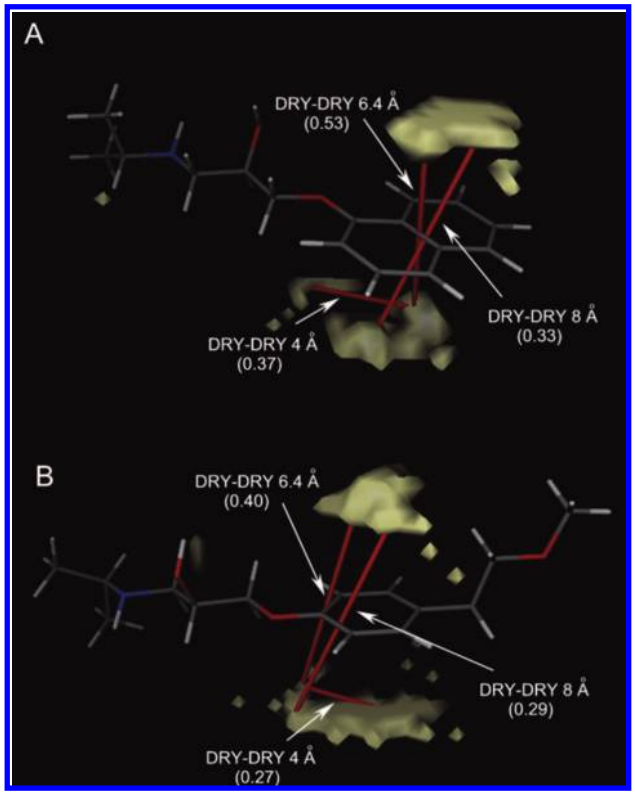

Figure 4. GRIND descriptors related to the presence of field generated by hydrophobic surfaces of molecules. DRY molecular interaction fields are shown for propranolol (A) and metoprolol (B). Lines connect DRY-DRY node couples at distances of 4.0, 6.4, and 8.0 Å. The values of the corresponding DRY-DRY GRIND descriptors are shown in brackets.

Supporting Information Table S2). However, interpretation of these comparisons needs to be done with caution because the models for CYP inhibition are built on completely different data sets and validated using different validation procedures. Moreover, the CYP inhibition activities were expressed in different ways, such as by K_i, $K_{i\ apparent}$, percentage of inhibition, or by binary values (i.e., activity/no activity), rather than by the IC_{50} values used herein, which further complicates comparisons of the models (Supporting Information Table S2).

In many reports CYP models were built on small data sets of structurally similar compounds.[9−11,15,16,22,34,56−58] Although such "local" models may be reasonably statistically valid, they are limited in scope and will likely fail to provide good predictions for compounds structurally different from those in the training set. On the other hand, the earlier QSAR and 3D-QSAR CYP models constructed on large proprietary data sets (i.e., hundreds to thousands of compounds tested for inhibition on an individual CYP isoform) usually had a categorical outcome, i.e. they classified compounds as inhibitors/noninhibitors or as strong/medium/weak inhibitors, but the cut-offs for the measured inhibition activities varied substantially between models (Supporting Information Table S2).[9,12,25,26,31,37,39,40,59,60] Our generalized CYP model was constructed on a large data set, and it provides predictions of inhibitory activities on a continuous scale.

Cross-validation was applied to validate the majority of the published CYP inhibition models, as shown in Supporting Information Table S2. However, Q^2 values may vary significantly depending on whether 3-fold, 5-fold, or "leave-one-observation-out" CV procedure is used. Thus, simply comparing Q^2 values of two models could potentially lead to a biased assessment. Nonetheless, the data in Supporting Information Table S2 demonstrate that Q^2 values for the majority of the individual CYP inhibition models do not exceed 0.60, whereas the Q^2 value for the present generalized CYP model is 0.66.

Another approach used in many of the published reports for estimating the prediction accuracy of CYP inhibition models was to calculate the percentage of the test set compounds predicted with an error less than 0.5 or 1.0 log units of the inhibition activity to the total number of the compounds in the test set. As seen from Supporting Information Table S2 the prediction accuracy of the test set compounds with the error less than 0.5−1.0 log units was according to this approach 67−89% for the reports where such estimates were earlier published. Our results from the

"leave-one-CYP-isoform-out" validation for the generalized CYP model showed that for the majority of the CYP isoforms (i.e., CYP2A5, CYP2A6, CYP2C19, CYP2C8, CYP2C9, CYP2D6, CYP3A4, CYP3A5, and CYP3A7) more than 67% of the test set predictions occurred with an error less than one log IC_{50} unit. However, it should be stressed that we predicted pIC_{50}s for completely new CYP isoforms and for new compounds, whereas for all earlier studies (Supporting Information Table S2) the predictions were for the same CYP isoform.

Validation of the present generalized CYP model on 10% of randomly excluded data showed that RMSEP 0.65−0.81 log units is to be expected when predicting new ligand data for any CYP isoform already included in the generalized model. The prediction error according to this validation procedure is thus comparable or lower than the RMSEP of the earlier published models (see Supporting Information Table S2).

The generalized CYP modeling approach presented in this study is not constrained to the use of GRIND descriptors for ligands, z-scales for CYPs, or PLS as a correlation technique. In fact, other types of descriptors might be used for the organic compounds and CYPs as well as any other suited machine learning techniques could be used for constructing the model. A main advantage of using GRIND descriptors is their alignment independence, however. Using z-scales for the CYP enzymes also has an advantage, as it allows simple quantitative description of CYP isoforms (e.g., from other species) and polymorphic variants from sequence only. This brings a possibility to predict inhibition of ligands to other CYP enzymes and CYP isoforms (Figure 2). However, the large number of ligand and CYP descriptors generated could potentially complicate modeling and lead to overfits. To decrease the risk we reduced the number of descriptors by preprocessing with PCA. Although PCA-based descriptors may be less interpretable, they contain original information and are of high value, e.g. in the case when one is interested in a fast computational filter for predicting *a priori* pIC_{50} activities for new CYP-ligand pairs. This is because the original PCA models can be used to form new PCA-based descriptors for new CYPs and new ligands. Although it may then happen that new entities fall outside the scope of a PCA model, this can be identified using established procedures.[41] In case a new CYP-ligand pair falls out of the PCA models' scope, pIC_{50} predictions for them should be treated with caution.

Many common human CYP polymorphisms have accumulated along racial lines, and it would be highly cost-effective to determine a drug candidate's pharmacogenetic profiles regarding these polymorphisms before it enters into clinical trial. Hence, the introduction of population CYP enzyme variability in the evaluation of ADMET characteristics for candidate drugs should be widely integrated in the early stages of drug development in order to reduce late expensive drug failures. The proteochemometric approach provides a general strategy for the analysis of a variety of CYP isoforms with multiple ligands. Our results suggest that the general proteochemometrics CYP model might be used as a computational filter for "*a priori*" predictions of potential drug-drug interactions in early as well as late stages of the development of drug candidates. This would require a proper description of the CYP isoforms according to their sequence in the same fashion that we described the other CYPs herein; then the model could be used to make predictions for their inhibition by candidate drugs. Our model also provides insight into the mechanisms underlying CYP inhibition, identifying molecular features critical to inhibition kinetics. The proteochemometric approach can assist in discovering robust inhibition factors and advance our understanding of the less-studied CYP P450 enzymes.

Furthermore, upon possessing information regarding the most important MIFs of new potential compounds (i.e., as shown in Table 2), it should be possible to "design out" potential CYP inhibition by introducing the proper chemical modifications in drug candidates. A proteochemometric model as presented herein could be used to screen virtual libraries of compounds, enable analysis of the multitude of complex cross-interactions governing CYP inhibition (i.e., in addition to the properties revealed in Table 2), and may aid in the improvement of metabolic properties of novel compounds in drug development. However, we propose that proteochemometrics should be applied on much larger proprietary chemical collections available within the pharmaceutical industry. This should result in broader generalized CYP inhibition models that cover a much larger interaction space for the CYPs and their inhibitors than covered by the model herein.

We envisage that models similar to the one presented in this study will be of great interest and practical utility for the pharmaceutical industry and will encourage further research in pharmacogenomics. The general approach described here should enable the study of other important polymorphic protein systems involved in absorption, distribution, metabolism, and elimination of xenobiotics, such as transporters and uridine 5′-triphosphate glucuronosyltransferases. The model could also enable the prediction of genetic subpopulations expected to respond adversely to a drug. From a theoretical point of view the proteochemometric approach is unlimited regarding the number of CYP subtypes and polymorphic variants that can be included in a model. Further development of generalized proteochemometric models of CYP is likely to offer a competitive advantage in drug development.

ACKNOWLEDGMENT

The research was supported by a grant from the Swedish Research Council (04X-05957).

Supporting Information Available: Data set used for construction of the generalized proteochemometric CYP model (Table S1), comparison of the generalized model for CYP inhibition with published models for inhibition of individual CYPs (Table S2), and alignment of 14 CYP enzymes used herein (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bu, H-Z. A Literature Review of Enzyme Kinetic Parameters for CYP3A4-mediated Metabolic Reactions of 113 Drugs in Human Liver Microsomes: Structure-kinetics Relationship Assessment. *Curr. Drug. Metab.* **2006**, *7*, 231–249.

(2) Guengerich, F. P. Cytochrome P450: What Have We Learned and What Are the Future Issues. *Drug Metab. Rev.* **2004**, *36*, 159–197.

(3) Wolf, C. R.; Smith, G.; Smith, R. L. Science, Medicine, and the Future: Pharmacogenetics. *Br. Med. J.* **2000**, *320*, 987–990.
(4) Arimoto, R. Computational Models for Predicting Interactions with Cytochrome P450 Enzyme. *Curr. Top. Med. Chem.* **2006**, *6*, 1609–1618.
(5) Doucet, J.; Jego, A.; Noel, D.; Geffroy, C. E.; Capet, C.; Coquard, A.; Couffin, E.; Fauchais, A. L.; Chassagne, P.; Mouton-Schleifer, D. Preventable and Non-preventable Risk Factors for Adverse Drug Events Related to Hospital Admissions in the Elderly: A Prospective Study. *Clin. Drug Invest.* **2002**, *22*, 385–392.
(6) Fuhr, U. Induction of Drug Metabolizing Enzymes: Pharmacokinetic and Toxicological Consequences in Humans. *Clin. Pharmacokinet.* **2000**, *38*, 493–504.
(7) Haining, R. L.; Yu, A. In *Drug Metabolizing Enzymes: Cytochrome P450 and Other Enzymes in Drug Discovery and Development*; Fisher, M., Lee, J., Obach, S., Eds.; FontisMedia: Lausanne, Switzerland, 2003; pp 375−415.
(8) Vermeulen, N. P. E. Prediction of Drug Metabolism: the Case of Cytochrome P450 2D6. *Curr. Top. Med. Chem.* **2003**, *3*, 1227–1239.
(9) Afzelius, L.; Masimirembwa, C. M.; Karlen, A.; Andersson, T. B.; Zamora, I. Discriminant and Quantitative PLS Analysis of Competitive CYP2C9 Inhibitors Versus Non-inhibitors Using Alignment Independent GRIND Descriptors. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 443–458.
(10) Afzelius, L.; Zamora, I.; Masimirembwa, C. M.; Karlen, A.; Andersson, T. B.; Mecucci, S.; Baroni, M.; Cruciani, G. Conformer- and Alignment-independent Model for Predicting Structurally Diverse Competitive CYP2C9 Inhibitors. *J. Med. Chem.* **2004**, *47*, 907–914.
(11) Afzelius, L.; Zamora, I.; Ridderstrom, M.; Andersson, T. B.; Karlen, A.; Masimirembwa, C. M. Competitive CYP2C9 Inhibitors: Enzyme Inhibition Studies, Protein Homology Modeling, and Three-dimensional Quantitative Structure-activity Relationship Analysis. *Mol. Pharmacol.* **2001**, *59*, 909–919.
(12) Arimoto, R.; Prasad, M-A.; Gifford, E. M. Development of CYP3A4 Inhibition Models: Comparisons of Machine-Learning Techniques and Molecular Descriptors. *J. Biomol. Screen.* **2005**, *10*, 197–205.
(13) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A Rapid Computational Filter for Cytochrome P450 1A2 Inhibition Potential of Compound Libraries. *J. Med. Chem.* **2005**, *48*, 5154–5161.
(14) Crivori, P.; Poggesi, I. Predictive Model for Identifying Potential CYP2D6 Inhibitors. *Pharmacol. Toxicol.* **2005**, *96*, 251–253.
(15) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three and Four Dimensional-quantitative Structure Activity Relationship (3D/4D-QSAR) Analyses of CYP2D6 Inhibitors. *Pharmacogenetics* **1999**, *9*, 477–489.
(16) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-dimensional Quantitative Structure Activity Relationship of Cytochrome P-450 3A4 Inhibitors. *J. Pharmacol. Exp. Ther.* **1999**, *290*, 429–438.
(17) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-dimensional Quantitative Structure Activity Relationship (3D/4D-QSAR) Analyses of CYP2C9 Inhibitors. *Drug Metab. Dispos.* **2002**, *28*, 994–1002.
(18) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmocophore and Three-dimensional Quantitative Structure Activity Relationship Methods for Modeling Cytochrome P450 Active Sites. *Drug Metab. Dispos.* **2001**, *29*, 936–944.
(19) Ekins, S.; Stresser, D. M.; Williams, J. A. In Vitro and Pharmacophore Insights Into CYP3A Enzymes. *Trends Pharmacol. Sci.* **2003**, *24*, 161–166.
(20) Hutzler, J. M.; Walker, G. S.; Wienkers, L. C. Inhibition of Cytochrome P450 2D6: Structure-activity Studies Using a Series of Quinidine and Quinine Analogues. *Chem. Res. Toxicol.* **2003**, *16*, 450–459.
(21) Jalaie, M.; Arimoto, R.; Gifford, E.; Schefzick, S.; Waller, C. L. Prediction of Drug-like Molecular Properties: Modeling Cytochrome P450 Interactions. *Methods Mol. Biol.* **2004**, *275*, 449–520.
(22) Jones, J. P.; He, M.; Trager, W. F.; Rettie, A. E. Three-dimensional Quantitative Structure-activity Relationship for Inhibitors of Cytochrome P4502C9. *Drug Metab. Dispos.* **1996**, *24*, 1–6.
(23) Kemp, C. A.; Flanagan, J. U.; van Eldrik, A. J.; Marechal, J-D.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. Validation of Model of Cytochrome P450 2D6: An in Silico Tool for Predicting Metabolism and Inhibition. *J. Med. Chem.* **2004**, *47*, 5340–5346.
(24) Korhonen, L. E.; Rahnasto, M.; Mahonen, N. J.; Wittekindt, C.; Poso, A.; Juvonen, R. O.; Raunio, H. Predictive Three-dimensioanl Quantitative Structure-activity Relationship of Cytochrome P450 1A2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 3808–3815.
(25) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. A Support Vector Machine Approach to Classify Human Cytochrome P450 3A4 Inhibitors. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 189–201.
(26) Kriegl, J. M.; Eriksson, L.; Arnhold, T.; Beck, B.; Johansson, E.; Fox, T. Multivariate Modeling of Cytochrome P450 3A4 Inhibition. *Eur. J. Pharm. Sci.* **2005**, *24*, 451–463.
(27) Lewis, D. F. V.; Dickins, M. Quantitative Structure-activity Relationships (QSARs) within Series of Inhibitors for Mammalian Cytochromes P450 (CYPs). *J. Enzyme Inhib.* **2001**, *16*, 321–330.
(28) Lewis, D. F. V.; Lake, B. G.; Dickins, M. Quantitative Structure-activity Relationships (QSARs) in CYP3A4 Inhibitors: The Importance of Lipophilic Character and Hydrogen Bonding. *J. Enzyme Inhib. Med. Chem.* **2006**, *21*, 127–132.
(29) Lewis, D. F. V.; Modi, S.; Dickins, M. Structure-activity Relationship for Human Cytochrome P450 Substrates and Inhibitors. *Drug Metab. Rev.* **2002**, *34*, 69–82.
(30) Marechal, J. D.; Yu, J.; Brown, S.; Kapelioukh, I.; Rankin, E. M.; Wolf, R.; Roberts, G. C.; Paine, M. J.; Sutcliffe, M. J. In Silico and in Vitro Screening for Inhibition of Cytochrome P450 CYP3A4 by Comedications Commonly Used by Patients with Cancer. *Drug Metab. Dispos.* **2006**, *34*, 534–538.
(31) Molnar, L.; Keseru, G. M. A Neural Network Based Virtual Screening of Cytochrome P450 3A4 Inhibitors. *Bioorg. Med. Chem. Lett.* **2002**, *12*, 419–421.
(32) Obach, R. S.; Walsky, R. L.; Venkatakrishnan, K.; Gaman, E. A.; Houston, J. B.; Tremaine, L. M. The Utility of in Vitro Cytochrome P450 Inhibition Data in the Prediction of Drug-drug Interactions. *J. Pharmacol. Exp. Ther.* **2006**, *316*, 336–348.
(33) Rahnasto, M.; Raunio, H.; Poso, A.; Wittekindt, C.; Juvonen, R. O. Quantitative Structure-activity Relationship Analysis of Inhibitors of the Nicotine Metabolizing CYP2A6 Enzyme. *J. Med. Chem.* **2005**, *48*, 440–449.
(34) Rao, S.; Aoyama, R.; Schrag, M.; Trager, W. F.; Rettie, A.; Jones, J. P. A Refined 3-dimensional QSAR of Cytochrome P450 2C9: Computational Predictions of Drug Interactions. *J. Med. Chem.* **2000**, *43*, 2789–2796.
(35) Riley, R. J.; Parker, A. J.; Trigg, S.; Manners, C. N. Development of a Generalized, Quantitative Physicochemical Model of CYP3A4 Inibition for Use in Early Drug Discovery. *Pharm. Res.* **2001**, *18*, 652–655.
(36) Strobl, G. R.; von Kruedener, S.; Stockigt, J.; Guengerich, F. P.; Wolff, T. Development of a Pharmacophore for Inhibition of Human Liver Cytochrome P-4502D6: Molecular Modeling and Inhibition Studies. *J. Med. Chem.* **1993**, *36*, 1136–1145.
(37) Susnow, R. G.; Dixon, S. L. Use of Robust Classification Techniques for the Prediction of Human Cytochrome P450 2D6 Inhibition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1308–1315.
(38) Wanchana, S.; Yamashita, F.; Hashida, M. QSAR Analysis of the Inhibition of Recombinant CYP3A4 Activity by Structurally Diverse Compounds Using a Genetic Algorithm-combined Partial Least Squares Method. *Pharm. Res. (N. Y.)* **2003**, *20*, 1401–1408.
(39) Yap, C. W.; Chen, Y. Z. Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.
(40) Zuegge, J.; Fechner, U.; Roche, O.; Parrott, N. J.; Engkvist, O.; Schneider, G. A Fast Virtual Screening Filter for Cytochrome P450 3A4 Inhibition Liability of Compound Libraries. *Quant. Struct.-Act. Relat.* **2002**, *21*, 249–256.
(41) Kontijevskis, A.; Petrovska, R.; Mutule, I.; Uhlen, S.; Komorowski, J.; Prusis, P.; Wikberg, J. E. S. Proteochemometric Analysis of Small Cyclic Peptides' Interaction with Wild-type and Chimeric Melanocortin Receptors. *Proteins* **2007**, *69*, 83–96.
(42) Kontijevskis, A.; Prusis, P.; Petrovska, R.; Yahorava, S.; Mutulis, F.; Mutule, I.; Komorowski, J.; Wikberg, J. E. S. A Look Inside HIV Resistance through Retroviral Protease Interaction Maps. *PLoS Comput. Biol.* **2007**, *3*, e48.
(43) Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteo-chemometrics Analysis of MSH Peptide Binding to Melanocortin Receptors. *Protein Eng.* **2002**, *15*, 305–311.
(44) Prusis, P.; Uhlen, S.; Petrovska, R.; Lapinsh, M.; Wikberg, J. E. S. Prediction of Indirect Interactions in Proteins. *BMC Bioinformatics* **2006**, *7*, 167.
(45) Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. Proteochemometrics Modeling of the Interaction of Amine G-protein Coupled Receptors with a Diverse Set of Ligands. *Mol. Pharmacol.* **2002**, *61*, 1465–1475.
(46) Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. S. QSAR and Proteo-chemometric Analysis of the Interaction of a Series of Organic Compounds with Melanocortin Receptor Subtypes. *J. Med. Chem.* **2003**, *46*, 2572–2579.
(47) Wikberg, J. E. S.; Lapinsh, M.; Prusis, P. Proteochemometrics: A tool for modeling the molecular interaction space. In *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*; Kubinyi, H.,

Müller, G., Mannhold, R., Folkers, G., Eds.; Weinheim: Wiley-VCH: 2004; pp 289−309.
(48) Wikberg, J.; Mutulis, F. Targeting melanocortin receptors: an approach to treat weight disorders and sexual dysfunction. *Nat. Rev. Drug Discovery* **2008**, *7*, 307–323.
(49) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.
(50) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. (2000) GRid-INdependent Descriptors (GRIND): A Novel Class of Alignment-independent Three-dimensional Molecular Descriptors. *J. Med. Chem.* **2000**, *43*, 3233–3243.
(51) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating Molecular Shape into the Alignment-free Grid-Independent Descriptors. *J. Med. Chem.* **2004**, *47*, 2805–2815.
(52) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
(53) Geladi, P.; Kowalski, B. R. Partial Least-squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
(54) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. PLS. In *Multi- and Megavariate Data Analysis Principles and Applications*; Eriksson, L.,Johansson, E., Kettaneh-Wold, N., Wold, S., Eds.; Umetrics Academy: Umeå, 2001; pp 71−112.
(55) Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nyström, Å.; Pettersen, J.; Bergman, R. Experimental Design and Optimization. *Chemom. Intell. Lab. Syst.* **1998**, *42*, 3–40.
(56) Poso, A.; Juvonen, R.; Gynther, J. Comparative Molecular Field Analysis of Compounds with CYP2A5 Binding Affinity. *Quant. Struct.-Act. Relat.* **1995**, *14*, 507–511.
(57) Poso, A.; Gynthera, J.; Juvonen, R. A Comparative Molecular Field Analysis of Cytochrome P450 2A5 and 2A6 Inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 195–202.
(58) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-Dimensional-Quantitative Structure Activity Relationship (3D/4D-QSAR) Analyses of CYP2C9 Inhibitors. *Drug Metab. Dispos.* **2000**, *28*, 994–1002.
(59) 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.
(60) Ekins, S.; Berbaum, J.; Harrison, R. K. Generation and Validation of Rapid Computational Filters for CYP2D6 and CYP3A4. *Drug Metab. Dispos.* **2003**, *31*, 1077–1080.

CI8000953