

## Recursive Partitioning for the Prediction of Cytochromes P450 2D6 and 1A2 Inhibition: Importance of the Quality of the Dataset

Julien Burton,<sup>\*,†</sup> Ismail Ijjaali,<sup>‡</sup> Olivier Barberan,<sup>‡</sup> François Petitet,<sup>‡</sup> Daniel P. Vercauteren,<sup>†</sup> and André Michel<sup>‡</sup>

Laboratoire de Physico-Chimie Informatique, Facultés Universitaires Notre-Dame de la Paix, 61 rue de Bruxelles, B-5000 Namur, Belgium, and Aureus-Pharma, 174 quai de Jemmapes, F-75010 Paris, France

Received March 9, 2006

The purpose of this study was to explore the use of detailed biological data in combination with a statistical learning method for predicting the CYP1A2 and CYP2D6 inhibition. Data were extracted from the Aureus–Pharma highly structured databases which contain precise measures and detailed experimental protocol concerning the inhibition of the two cytochromes. The methodology used was Recursive Partitioning, an easy and quick method to implement. The building of models was preceded by the evaluation of the chemical space covered by the datasets. The descriptors used are available in the MOE software suite. The models reached at least 80% of Accuracy and often exceeded this percentage for the Sensitivity (Recall), Specificity, and Precision parameters. CYP2D6 datasets provided 11 models with Accuracy over 80%, while CYP1A2 datasets counted 5 high-accuracy models. Our models can be useful to predict the ADME properties during the drug discovery process and are indicated for high-throughput screening.

### Introduction

One of the major reasons of failure in drug discovery projects is related to the poor pharmacokinetic and ADME (absorption, distribution, metabolism, and excretion) properties of drug candidates.<sup>1</sup> Cytochromes P450 (CYP<sup>a</sup>) play a crucial role in metabolism.<sup>2</sup> The proportion of drugs metabolized in human liver by CYP2D6 and CYP1A2 represents 19% and 10%, respectively, of the whole activity range of all cytochromes.<sup>3</sup> These percentages make them targets of choice for studying very early the potential inhibition by drug candidates. Forecasting inhibition of CYP2D6 and CYP1A2 can help to predict and manage drug–drug interactions (DDI).<sup>4,5</sup> It is thus particularly important to take into account the biotransformation and elimination of drugs during their development to determine as early as possible their potential interactions with CYPs. In this sense, predictive *in silico* models are very useful in ADME/DDI predictions.<sup>6,7</sup> However, in the case of CYPs, they are particularly challenging to build due to the relative low specificity of these enzymes. Compounds metabolized by cytochromes P450 are, indeed, structurally very dissimilar.

*In silico* studies, exploiting different methods as QSAR, pharmacophore modeling, molecular docking, etc., have already been published, and predicting CYP inhibition is more and more efficient.<sup>8–15</sup> The method chosen in our work is recursive partitioning (RP), which is known to be fast and which leads to easily interpretable results. RP is based on decision trees and has been used in diagnostic<sup>16</sup> as well as in high-throughput virtual screening.<sup>17–21</sup> More precisely, RP involves the creation of a decision tree composed of binary split nodes that divide the initial training set into smaller sets of higher purity, i.e., in

this study, sets containing a majority of inhibitors or a majority of noninhibitors. Each split node can be compared to a binary question (yes/no) regarding the value of a particular descriptor. After the creation of the tree, any other new compound, for which the descriptors used in the split nodes have been calculated, can be classified into an inhibitor or a noninhibitor type category; binary trees can thus be used to predict ADME properties. RP is known to be sensitive to the descriptors used, to unbalanced training sets, constituted for example with too many inhibitors, and to the composition of the datasets, that can radically change the decision tree.<sup>8</sup> In this paper, we focused on building efficient models for the prediction of CYP2D6 and CYP1A2 inhibition. Different parameters have been optimized and different datasets were constituted in order to propose predictive models as accurate as possible.

One of the main problems consists of having access to a sufficient number of structured data to build efficient models. That is why our approach will be based on three main characteristics of our datasets. The quantity of data had to be always sufficient (~100 compounds or more); the diversity of the datasets needs to cover a large chemical space; and different biological parameters linked to the compounds must be analyzed. Such a collection of data with these particular qualities differentiates ours from all other previous works.<sup>8,11–15</sup> The main purpose of this study is thus, with a simple methodology (RP), to investigate how structured biological data can help to constitute the most efficient training sets to build prediction models. The quantity, the quality, and the diversity of the datasets have been prone to discussion. Accordingly, the goal followed here is not to make a comparison between different methods to improve the models but to probe different parameters characterizing the training sets.

### Results and Discussion

**1. Datasets.** To achieve this study, two global datasets of 498 and 306 compounds were at disposal for CYP2D6 and CYP1A2, respectively. The content of each dataset is reported in Table 1, including the class thresholds as well as the number of inhibitors and noninhibitors in the sets.

\* To whom correspondence should be addressed. Phone: +32 (0)81 72 54 62. Fax: +32 (0)81 72 54 66. E-mail: julien.burton@fundp.ac.be.

<sup>†</sup> Facultés Universitaires Notre-Dame de la Paix.

<sup>‡</sup> Aureus-Pharma.

<sup>a</sup> Abbreviations: ADME, absorption distribution metabolism excretion; CYP, cytochrome P450; DDI, drug–drug interactions; QSAR, quantitative structure–activity relationship; RP, recursive partitioning; VSA, van der Waals surface area; AMMC, 3-[2-(*N,N*-diethyl-*N*-methylammonium)ethyl]-7-methoxy-4-methylcoumarin; MOE, molecular operating environment; PCA, principal component analysis; PC, principal component; SVM, support vector machine.

**Table 1.** Distribution of the Compounds in the Inhibitor and Noninhibitor Classes According to the Different Thresholds

dataset	class threshold, $\mu\text{M}$	inhibitors	noninhibitors
CYP2D6			
global	10	206	292
	25	247	251
bufuralol	3-30	74	64
dextromethorphan	3-30	86	48
AMMC	3-30	61	28
$K_i$	10	78	85
	3-30	51	51
$\text{IC}_{50}$	10	71	174
	3-30	47	142
CYP1A2			
global	50	154	152
$K_i$	30	41	40
$\text{IC}_{50}$	30	92	133

The chemical diversity of the training and test sets was assessed using a nearest neighbor searching algorithm as implemented in ChemAxon's application Compr.<sup>22</sup> In this algorithm, a weighted Euclidean distance calculation applies the Tanimoto (Jaccard) coefficient based on ChemAxon CF fingerprints. The dissimilarity between molecules is then given by:

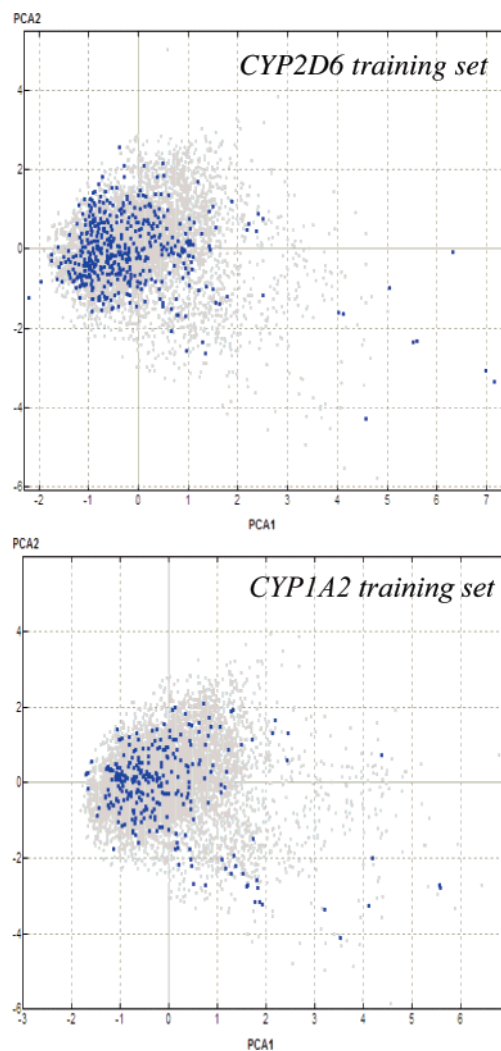
$$D(A,B) = 1 - T(A,B) = \{[1 - T(A,B)] + w_1[C_1(A) - C_1(B)]^2 + w_2[C_2(A) - C_2(B)]^2 + \dots\}^{1/2}$$

where  $w_1, w_2 \dots$  are weights,  $T(A,B)$ , the Tanimoto coefficient for molecules A and B, and  $C_i(A)$ , the value of descriptor  $i$  of molecule A.

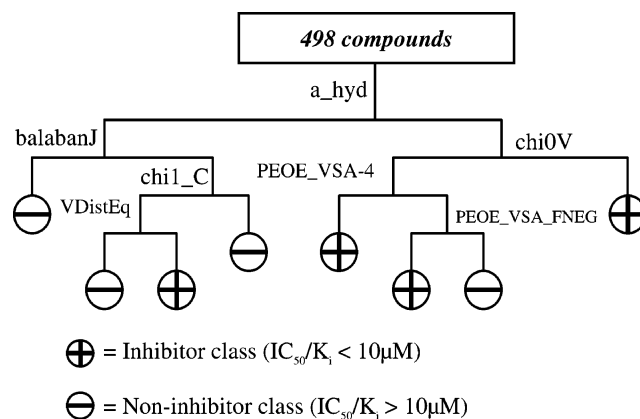
In the case of CYP2D6, the average and maximum nearest neighbor self-dissimilarities within the global training set were 68.6% and 99.6%, respectively. For the CYP1A2 data, values of the same range were obtained for the global training set, i.e., 68.8% and 96.6%. The obtained values suggest that the used datasets are sufficiently diverse. For both cytochromes, the same analysis was done on corresponding external sets. Results are discussed in Section 6.

To visualize the diversity of our datasets, molecules of the training sets were compared with a general database of compounds related to ADME measures found in the literature, i.e., with the AurSCOPE ADME/DDI database. The comparison was based on the two first Principal Components calculated from 32 P\_VSA descriptors of all the molecules. The well-distributed data for both 2D6 and 1A2 sets reinforce the idea that our datasets cover a large chemical space (Figure 1).

**2.1. CYP2D6. Global Dataset.** First, a global study was performed with 498 compounds related to CYP2D6 inhibition measures with mixed  $\text{IC}_{50}$  and  $K_i$  using 2D + P\_VSA descriptors. The accuracy of the different trees built with the different parameters, i.e., varying node split, depth, threshold of inhibitor/noninhibitor, allowed us to conclude that the default MOE parameters, node split = 10 and depth = 10, were adapted to our kind of study. Indeed, other parameter values did not radically change the overall Accuracy of the trees that always varied around 75%. The best model with this global dataset had an Accuracy of 78% with an inhibitors/noninhibitors cutoff of 10  $\mu\text{M}$ . Actually, this value could correspond to the expected hepatic blood concentration of typical drug-like molecule when administrated at therapeutic doses.<sup>8</sup> This model correctly predicted 168 out of the 206 inhibitors, with 82% of Sensitivity, and 219 of the 292 noninhibitors, with 75% of Specificity; the Precision associated to inhibitor compounds was 70%. The structure of the obtained tree with the descriptors involved is presented in Figure 2.



**Figure 1.** Data distribution of the CYP2D6 and CYP1A2 training sets (blue dots) compared to AUREUS PHARMA's AurSCOPE ADME/DDI database, release June 2005 (grey dots). The comparison is based on the two first principal components calculated from 32 P\_VSA descriptors.



**Figure 2.** Decision tree and its descriptors built with 498 compounds related to  $\text{IC}_{50}$  or  $K_i$  measures for CYP2D6 inhibition. 2D descriptors, including P\_VSA descriptors, were used, and the limit between inhibitors and noninhibitors was fixed to 10  $\mu\text{M}$ .

**2.2. CYP2D6. Probe Substrates Datasets.** To improve our models, more precise studies were performed based on the probe substrate with which the inhibition measures were made. Three main substrates were isolated to build consistent datasets, i.e., containing at least 80 compounds. These were bufuralol,

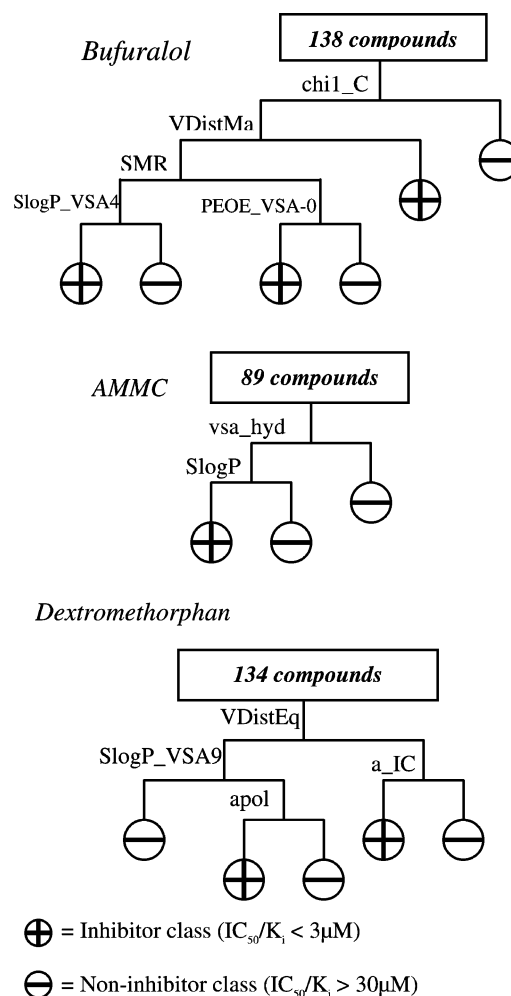
**Table 2.** Performance Parameters, Accuracy, Sensitivity, Specificity, and Precision (in %), for the Three Models Corresponding to Three CYP2D6 Probe Substrates Used in the Inhibition Measurement Protocols, Bufuralol, Dextromethorphan, and AMMC

	Accuracy	Sensitivity	Specificity	Precision
bufuralol	88	80	98	98
dextromethorphan	82	81	83	90
AMMC	89	82	92	82

dextromethorphan, and AMMC (3-[2-(*N,N*-diethyl-*N*-methylammonium)ethyl]-7-methoxy-4-methylcoumarin). Both sets of 2D and P\_VSA descriptors were used together, and new cutoff values for the classes were fixed, as the 10  $\mu\text{M}$  threshold did not provide good results (data not shown). Inhibitors were defined as compounds with  $K_i$  or  $\text{IC}_{50} < 3 \mu\text{M}$  and noninhibitors  $> 30 \mu\text{M}$ . Compounds between 3 and 30  $\mu\text{M}$  were removed, but the inhibitors/noninhibitors classes were more discriminating than the 10  $\mu\text{M}$  cutoff value. The 3 and 30  $\mu\text{M}$  cutoffs were also chosen because of the good repartition of the compounds in the two classes. Fixing these limits permitted us to keep enough molecules in each class. Moreover, other thresholds were tested in this particular case and did not provide results as good as the 3–30  $\mu\text{M}$  one. Generating a tree based on these sets led to the construction of three models; performance parameters as defined in Materials and Methods are shown in Table 2 for each of them. All the parameters are above 80%, but the highlight of the models is the particularly high Specificity, and Precision, i.e., 98%, of the bufuralol-based model. This last decision tree predicted 60 compounds to be inhibitors; only one of them was actually a noninhibitor. Observing and comparing the three trees, we conclude that they have a relatively small depth: 4, 3, 2 for bufuralol, dextromethorphan, and AMMC, respectively. Descriptors used in their nodes are significantly different as shown in Figure 3. No significant similarity between these trees can be reported despite their close performances.

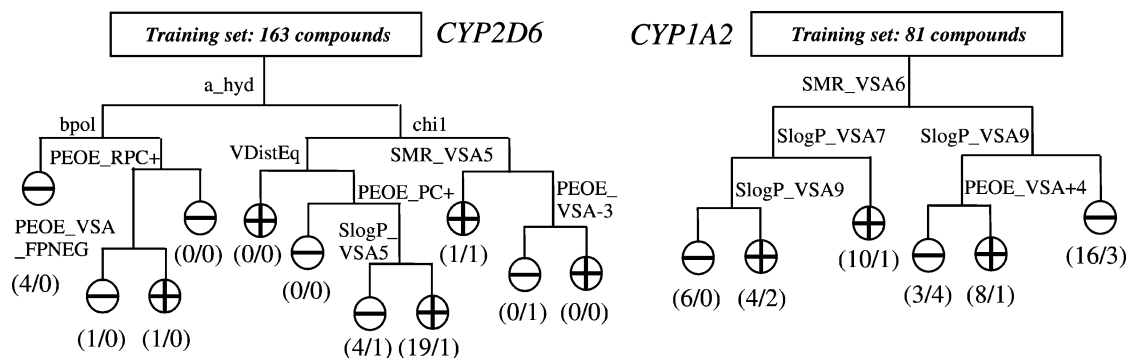
**2.3. CYP2D6. Separating  $K_i$  and  $\text{IC}_{50}$ .** Discriminating between the  $K_i$  and  $\text{IC}_{50}$  values could enhance the accuracy of CYP2D6 models. It could not be done for the substrate study due to the relative small size of the datasets. The distinction could be done for the global set and led to the constitution of a set of 163  $K_i$  compounds and a set of 245  $\text{IC}_{50}$  compounds. The  $K_i$  set provided the best results. A first model was generated with the 2D and P\_VSA descriptors together and a 10  $\mu\text{M}$  threshold which separated the dataset into 78 inhibitors and 85 noninhibitors. Its Accuracy equals 90%, Sensitivity, 88%, Specificity, 92%, and Precision, 90%, that constituted a really efficient model (Figure 4). Using the set of P\_VSA descriptors alone with 3–30  $\mu\text{M}$  thresholds led to quite similar model performances with an Accuracy of 90%, Sensitivity of 90%, Specificity of 91%, and Precision of 92%. Obviously, the trees cannot be compared because they are composed of totally different descriptors. The one built with the 2D + P\_VSA descriptors needed a depth of 5 and 10 nodes while the one based on P\_VSA alone has only a depth of 3 and is composed of 5 nodes (data not shown). In this last case, the P\_VSA descriptors are more “efficient”, as they needed a smaller tree to provide results as good as those with the 2D + P\_VSA descriptors. An interesting event that can be highlighted is that P\_VSA descriptors were taken into account in the building of the two trees; but mixed with pure 2D descriptors, the model obtained was more complex but equal in performances. Doubtless, this is due to “noise” generated by using too many descriptors in this particular situation.

Such good results encouraged us to investigate more deeply the  $K_i$  set with a multiclass study. Three classes were therefore



**Figure 3.** Decision trees and their descriptors built from three datasets corresponding to the three most frequent probe substrates (bufuralol, dextromethorphan, and AMMC) used in the inhibition experiments for CYP2D6.

defined. Compounds with  $K_i < 3 \mu\text{M}$  were identified as high inhibitors, that represented 51 molecules. Compounds with  $K_i$  between 3  $\mu\text{M}$  and 30  $\mu\text{M}$  were called medium inhibitors, with 65 molecules, and compounds with  $K_i > 30 \mu\text{M}$  were called poor inhibitors, with 47 molecules. This time, the combination of the 2D and P\_VSA descriptors led to the best model; the classification percentages of the training set for each class are presented in Table 3. The overall Accuracy of this model is 83%, and the Precision associated to each of the class is 85%, 82%, and 83%, for the high, medium, and poor inhibitors, respectively. The model correctly discriminates high and poor inhibitors, as when predicting high inhibitors, only 4% are taken as poor and when predicting poor inhibitors, none of the compounds are placed in the high inhibitors class. In the case of  $\text{IC}_{50}$  measures, several combinations between the descriptors sets and tree parameters were tested but none led to effective models. The key was actually to take into account the protonation stage of each molecule. Protonation of the basic functions, deprotonation of the acid ones, and a recalculation of the descriptors based this time on charged atoms succeeded in two more efficient trees. The effect of this manipulation was to enhance the electrostatic properties of the compounds. Actually, electrostatic properties are one of the key factors governing CYP2D6 inhibition (cf. Section 7. Comparison with Earlier Studies). Their inhibitors/noninhibitors classes were defined as



**Figure 4.** Two of the best models from the CYP2D6 ( $K_i$  dataset, 10  $\mu\text{M}$  threshold, 2D and P\_VSA descriptors) and CYP1A2 ( $K_i$  dataset, 30  $\mu\text{M}$  threshold, P\_VSA descriptors) datasets. Both models have been validated with an external dataset of 34 and 58 compounds for CYP2D6 and CYP1A2, respectively. + and - signs mean that the class assigned to a leaf is inhibitor or noninhibitor, respectively. The distribution of molecules of the test set in each leaf is positioned between brackets (number of correctly classified compounds/number of misclassified compounds).

**Table 3.** Prediction Percentages for the Multiclass Study Based on the  $K_i$  Measures from the CYP2D6 Inhibition Experiments<sup>a</sup>

class	prediction		
	high, %	medium, %	poor, %
high	<b>78</b>	18	4
medium	11	<b>78</b>	11
poor	0	4	<b>96</b>

<sup>a</sup> Percentages of true prediction are noted in bold. Three classes were considered: high inhibitors ( $K_i < 3\mu\text{M}$ ), medium inhibitors ( $3\mu\text{M} < K_i < 30\mu\text{M}$ ), and poor inhibitors ( $K_i > 30\mu\text{M}$ ).

< or > 10  $\mu\text{M}$  for the first one and <3 and >30  $\mu\text{M}$  for the other; they both presented an overall Accuracy of 85%. The 10  $\mu\text{M}$  threshold tree presented 89% of Sensitivity, 84% of Specificity, and 69% of Precision. The respective parameters for the second tree are 83%, 89%, and 72%, respectively. Precision values are somewhat disappointing in regard of the previous values obtained for the  $K_i$  or probe substrate based models. That can be explained by the unbalanced  $\text{IC}_{50}$  datasets. Indeed, the sets contained about three-quarters of noninhibitor compounds. A poor classification for these has a worse influence on the Precision than for a balanced dataset, considering that the proportion of false positives is larger. The two  $\text{IC}_{50}$  trees are rather structurally different as the first one, with the 10  $\mu\text{M}$  threshold, contained only 2 nodes for a depth of 2 and the other one had 4 nodes and a depth of 4. Comparing the  $\text{IC}_{50}$  and  $K_i$  models, it is obvious that the  $K_i$ -based models are more reliable than the  $\text{IC}_{50}$ -based models, as the difference of Accuracy is 5% in favor of  $K_i$ , i.e., 90% for both the  $K_i$  models and 85% for both the  $\text{IC}_{50}$  models. This can be explained by the fact that  $K_i$  values are intrinsic constants, whereas  $\text{IC}_{50}$  values are extrinsic constants.  $\text{IC}_{50}$  values, in contrast to  $K_i$  values, are dependent on the type of substrate, the concentration of substrate, and incubation conditions (protein concentration or incubation times, etc.). As a consequence, better classifications could be obtained when  $K_i$  values are used instead of  $\text{IC}_{50}$  values.

**3.1. CYP1A2. Defining a Threshold.** The good quality results obtained for CYP2D6 encouraged us to apply our working method to another interesting cytochrome P450. The CYP1A2 dataset was composed of 225  $\text{IC}_{50}$  values and 81  $K_i$  values. Literature sources about CYP1A2 did not provide a precise inhibitor/noninhibitor threshold such as the 10  $\mu\text{M}$  value for CYP2D6. Therefore, different trees were built with different thresholds to determine the best limit between the two classes. Two types of descriptors were tested for this study: 2D descriptors and P\_VSA descriptors. We analyzed different thresholds between 3  $\mu\text{M}$  and 100  $\mu\text{M}$  and concluded that the optimum

results were obtained for thresholds between 30 and 50  $\mu\text{M}$ . It is precisely the same range of threshold used in the recently published study by Chohan et al.<sup>14</sup>

**3.2. CYP1A2. Separating  $K_i$  and  $\text{IC}_{50}$ .** As shown in the CYP2D6 study, separating  $\text{IC}_{50}$  and  $K_i$  generated good prediction models. The  $\text{IC}_{50}$  dataset made of 225 molecules was used and, combined with the set of P\_VSA descriptors and a 30  $\mu\text{M}$  class limit, gave an 86% Accuracy model. Sensitivity, Specificity, and Precision were of 83%, 88%, and 83%, respectively. With the same descriptors and the same threshold, the  $K_i$  dataset of 81 molecules led to an overall Accuracy of 89%, Sensitivity of 95%, Specificity of 83%, and Precision of 85% (Figure 4). 39 of the 41 inhibitors were correctly classified. As for the CYP2D6 models, predicting classes based on the  $K_i$  values is more efficient. The 30- $\mu\text{M}$  threshold separates once again the  $K_i$  dataset in two equal parts, i.e., 41 inhibitors and 40 noninhibitors, following the example of the 50  $\mu\text{M}$  threshold on the global set. If 2D descriptors are added to the P\_VSA ones, the  $K_i$ -based model's performances are quite similar with an overall Accuracy of 89%, Sensitivity of 86%, Specificity of 92%, and Precision of 93%.

**4. Using 3D Descriptors.** To further improve our models, 3D descriptors were also exploited. For CYP2D6, the use of the set of 3D descriptors allowed us to build a model with an 84% Accuracy, 83% Sensitivity, 85% Specificity, and 85% Precision. In a rather intuitive way, the addition of 2D descriptors to the 3D ones allowed improvement of the model. Accuracy, Sensitivity, Specificity, and Precision, were 87%, 88%, 87%, and 87% respectively. With a set of mixed 2D and 3D descriptors, the trees gained 2 to 5% on each accuracy parameter compared to 3D descriptors used alone.

For the global CYP1A2 dataset, a tree was built based on classes delimited by a 50  $\mu\text{M}$  threshold. The Accuracy was 90%, Sensitivity, 91%, Specificity, 88%, and Precision, 89%. A multiclass study was also performed. Very discriminating thresholds were used: high inhibitors were under 10  $\mu\text{M}$ , medium between 10 and 200  $\mu\text{M}$  and poor over 200  $\mu\text{M}$ . The best results were obtained with a combination of all the descriptors, 2D, P\_VSA, and 3D. Actually, these values split the dataset into three reasonably equivalent groups of 100, 80, and 108 molecules, respectively. The overall Accuracy of the obtained model is 86% and the Precision for the high, medium, and poor inhibitors is 83%, 85%, and 90%, respectively. Classification results for all the classes are presented in Table 4. The detailed structure of this tree is reported in Figure 5. The tree (depth of 7, 27 nodes) resulting from the use of the three classes that lead to three types of leaves is rather complex.

**Table 4.** Prediction Percentages for the Multiclass Study Mixing the 2D, P\_VSA, and 3D Descriptors for the CYP1A2 Inhibition Experiments<sup>a</sup>

class	prediction		
	high, %	medium, %	poor, %
high	<b>89</b>	5	6
medium	16	<b>78</b>	6
poor	5	6	<b>89</b>

<sup>a</sup> Percentages of true prediction are noted in bold. High inhibitors are compounds whose  $IC_{50}$  or  $K_i$  are under  $10 \mu M$ , medium inhibitors are between  $10 \mu M$  and  $200 \mu M$ , and poor inhibitors are over  $200 \mu M$ .

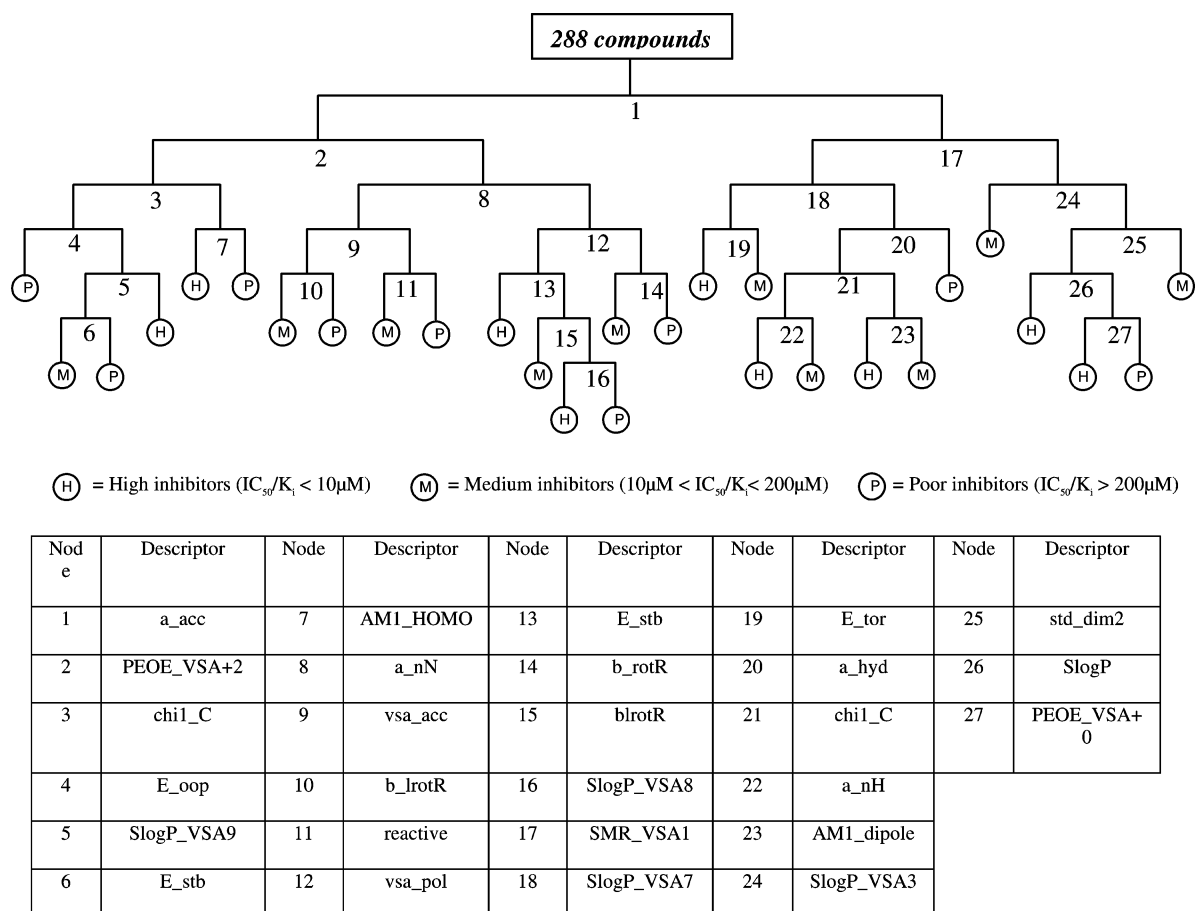
**5. Pertinent Descriptors.** Principal component analysis (PCA) reduces the dimensions of measured variables, i.e., our descriptors, to the representative principal components (PCs). As the set of PCs is smaller than the one of descriptors, a PC explains a greater variability of the dataset than a single descriptor. That kind of investigation was executed on the global datasets for CYP2D6 and CYP1A2, focusing only on 2D and P\_VSA descriptors. The first two PCs did already describe about 40 and 10% of the whole variability for CYP2D6 and 45 and 10% for CYP1A2. As PCs are linear combinations of the descriptors, searching for descriptors with a high coefficient in the combination led to the isolation of the most influential ones. In doing so, 39 descriptors were noted for the two first PCs of CYP2D6 and CYP1A2 datasets. A majority of these descriptors could be found in the four PCs considered. Very traditional descriptors, such as density, SlogP, number of double/triple/rotatable bonds, number of halogen atoms, and number of H-bonds donor/acceptor, were defined as pertinent. Several

models based on this set of 39 descriptors were built but did not led to better results compared to the ones obtained with the previous trees.

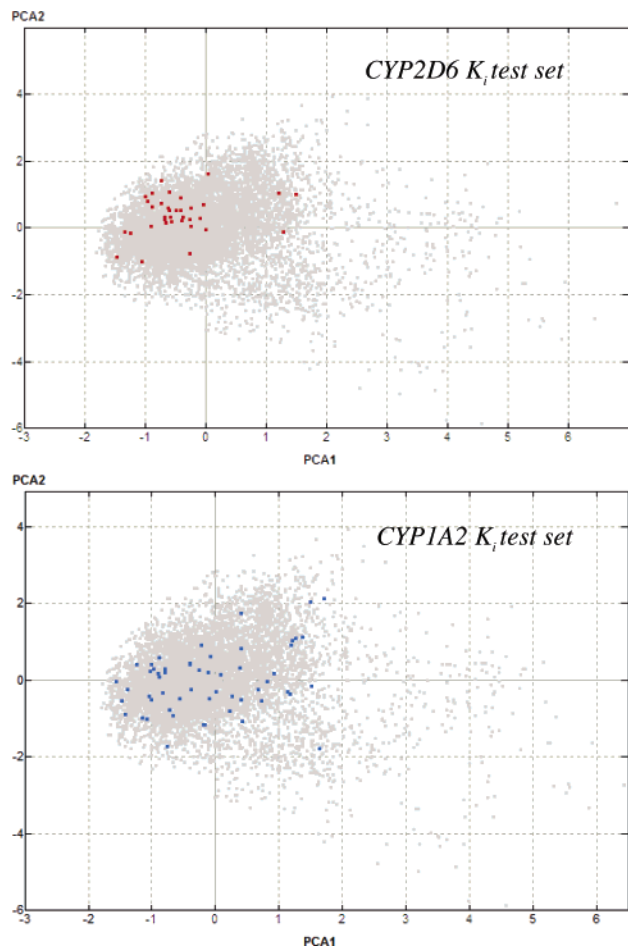
For several molecules, 18 of all the 3D descriptors could not be calculated. Therefore, 13 molecules from the CYP1A2 dataset and 15 from the CYP2D6 dataset were removed to perform the 3D studies. When analyzing the descriptors of the 15 models we retained, it can be concluded that some descriptors are more used than others. SlogP appears 5 times and P\_VSA descriptors based on SlogP (named SlogP\_VSAx) 16 times. Another 2D descriptor is also counted 5 times, VDistEq, which is based on the distance matrix of the molecule. For 3D descriptors, the ones calculated from the AM1 method are the most frequent with 10 occurrences. E\_Tor (torsion potential energy) is counted 5 times. A deeper discussion about controlling factors for the inhibition of CYP1A2 and CYP2D6 is presented in Section 7.

**6. External Validation.** To confirm the performance of our models, two external test sets of 34 and 58 molecules related to the CYP2D6 and CYP1A2  $K_i$  experiments were collected. Their diversity was computed exactly as for the previous training sets. For the CYP2D6 external test set, the average and maximum nearest neighbor dissimilarities were 60.4% and 86.2%. When considering the diversity of the test set versus the corresponding  $K_i$  training set, the average and maximum dissimilarities were 52.9% and 79.7%, respectively. For CYP1A2, these values were 69.9% and 94.5% for the self-dissimilarity evaluation of the external set, 80.4% and 98.7% for the test/training sets dissimilarity comparison.

For both cytochromes, Figure 6 shows a good coverage of chemical space. The molecules were submitted to the model



**Figure 5.** Decision tree built for a multiclass analysis for the CYP1A2 inhibition. Three classes of inhibitors were used: high inhibitors ( $K_i$  or  $IC_{50} < 10 \mu M$ ), medium inhibitors ( $10 \mu M < K_i$  or  $IC_{50} < 200 \mu M$ ), and poor inhibitors ( $K_i$  or  $IC_{50} > 200 \mu M$ ). Each node is labeled by a number that corresponds to an entry of the table. The table contains the descriptor used in each node.



**Figure 6.** Data distribution of CYP2D6 and CYP1A2  $K_i$  test sets (red and blue dots, respectively) compared to Aureus Pharma's AurSCOPE ADME/DDI database, release June 2005 (grey dots). The comparison is based on the two first principal components calculated from 32 P\_VSA descriptors.

we presented as the best: CYP2D6  $K_i$ -based tree built with 2D and P\_VSA descriptors (Figure 4). This validation was successful, as only two false negatives and two false positives were detected. The parameters of this classification are close from those obtained with the training set, as the Accuracy is 89% (90% for the training set), Sensitivity is 91% (88%), Specificity is 81% (92%), and Precision is 91% (90%). The detail of the repartition of the test set compounds in different leaves shows that one leaf is particularly populated with 19 inhibitors and one noninhibitor.

For CYP1A2, the  $K_i$ -based model built using VSA descriptors with 30  $\mu\text{M}$  as threshold and best Sensitivity was evaluated on the test set. Despite the larger dissimilarity between the training and test sets, very reasonable parameters were obtained since Accuracy is 81% (89% for the training set), Sensitivity is 76% (95%), Specificity is 86% (83%), and Precision is 85% (85%). Here, the distribution of the test set compounds is more widespread; the validation of the model is thus reinforced.

Correlation matrixes of the main molecular descriptors involved in the classification are given for the best 2D6 and 1A2 models in the Supporting Information; they indicate a low degree of correlation among them.

In general the models are likely to correctly predict the test set compounds when we consider similar datasets. Here, the inter-dissimilarity between each of the training sets and its corresponding test set along with obtained predictions reflects the global quality of the built models.

**7. Comparison with Earlier Studies.** Several papers have already described various pharmacophore models<sup>23–25</sup> as well as some 3D QSAR studies regarding CYP2D6 and CYP1A2.<sup>10,12,26,27</sup> However, it is known that such approaches are not convenient when large and diverse chemical and biological datasets are available from different sources. The use of machine learning techniques with categorical data is consequently gaining popularity and several studies on cytochromes P450 have been reported due to the ability of using such models to screen rapidly large molecular libraries (Table 5). Ekins et al.<sup>28</sup> employed RP to model the percentage of inhibition of CYP2D6 using a large dataset of 1759 molecules in combination with over 2500 augmented atom descriptors. Models were tested on 98 external molecules leading to Spearman's value of 0.61 with 50 compounds correctly predicted (51%). It should be noted that, contrary to the present study, biological activity was expressed in terms of percentage of inhibition, which is less reliable than  $\text{IC}_{50}$  or  $K_i$  values. Similarly to Ekins et al., consensus recursive partitioning was used by Susnow et al.<sup>8</sup> to identify inhibitors of CYP2D6. These authors used 25 in-house 2D molecular descriptors computed for a training set of 100 compounds. Internal validation tests indicated an overall classification of 75%. When applied to a 51 molecule external set assembled from literature, the model led to an Accuracy of 100% for 10 inhibitors and 75% for 41 noninhibitors. Recently, O'Brien and de Groot<sup>13</sup> used other machine learning methods including neural networks and Bayesian models; a consensus model combining these methods predicted 87% of positives and 75% of negatives. Although these three studies brought an efficient overall prediction of CYP2D6 inhibitors, no precise information on the selected descriptors and their interpretation was reported.

Yap and Chen<sup>15</sup> explored the use of the support vector machine (SVM) for predicting inhibition for CYP3A4, CYP2C9, and CYP2D6. The 2D6 training and validation datasets consisted of 602 and 198 molecules, respectively. These authors used 1607 structural and physicochemical descriptors to compute the average similarity value between all pairs of compounds in the dataset in a similar way to our method for analyzing the chemical diversity of the training and validation sets. Descriptors encoding electrostatic and hydrophobic characteristics were selected as relevant descriptors to classify inhibitors and noninhibitors of CYP2D6. Our findings are consistent with these earlier studies. When considering the model we presented as the best, i.e.,  $K_i$ -based tree (Figure 4), the descriptors selected on first nodes were a\_hyd, number of hydrophobic atoms, chi1, Hall atomic connectivity index,<sup>29,30</sup> bpol, sum of the absolute value of atomic polarizabilities of all bonded atoms in the molecule, PEOE\_RPC+, relative positive partial charge, and the SMR\_VSA5 descriptor. The average values of these descriptors for inhibitors and noninhibitors of 2D6 dataset are gathered in Table 6. These descriptors indicate the hydrophobicity, shape, and electrostatic contributions, as suggested by pharmacophoric modeling of inhibitors of CYP2D6.<sup>31</sup> A comparison between these descriptors for some similar molecules belonging to different classes is given in Table 7. Hydroxynefazodone and nefazodone are both 2D6 inhibitors but differently predicted. This is due to the corresponding values of the size related descriptor SMR\_VSA5, 177.5 and 196.52 for hydroxynefazodone and nefazodone, respectively. Also interestingly, clomipramine and imipramine, which differ by a Cl substituent, belong to different classes and have distinct SMR\_VSA5 values while the other relevant descriptors are very close. This information can be used to propose different substituents around the

**Table 5.** Summary and Comparison of Some Published Machine Learning Studies for 2D6 and 1A2 Cytochromes

CYP	techniques	training dataset	external dataset	Biological activity	used descriptors	performance
2D6 <sup>28</sup>	recursive partitioning	1759 from commercial database	98 from commercial source	% of inhibition	2500 commercial 2D descriptors	$r^2 = 0.88$ (training set)
2D6 <sup>8</sup>	ensemble recursive partitioning	100 from literature	51 from literature	$K_i$	in-house 2D descriptors	Spearman's $\rho = 0.61$ (test set) Accuracy = 100% (10 inhibitors) Accuracy = 75% (41 noninhibitors)
2D6 <sup>13</sup>	neural network	1810 from commercial source	600 from commercial source	$IC_{50}$	2D descriptors	Sensitivity: 86%, 83%
2D6 <sup>15</sup>	Bayesian model SVM	602 from literature and commercial sources	100 from literature	$K_i$ and others not specified	1607 2D and 3D commercial descriptors	Specificity: 84%, 80% Sensitivity = 75%
1A2 <sup>14</sup>	PLS MLR CART BNN	109 (22 in-house, 87 from commercial source)	68 from commercial source	$IC_{50}$ , $K_i$	in-house 2D descriptors	Specificity = 96.3% $r^2 = 0.72$ (training set) $r^2 = 0.71$ (training set) $r^2 = 0.84$ (training set) $r^2 = 0.72$ (training set)



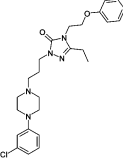
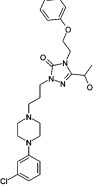
**Table 6.** Differences in the Values of the Selected Descriptors for CYP2D6 and CYP1A2 Inhibitors Classification

descriptor	CYP2D6		CYP1A2		
	average value		descriptor	average value	
	inhibitor	noninhibitor		inhibitor	noninhibitor
A_hydr	18.0	14.2	SMR_VSA6	15.1	56.2
bpol	30.4	25.9	SlogP_VSA7	124.6	78.4
Chi1	11.8	9.7	SlogP_VSA9	63.3	79.4
VdistEq	3.3	3.0	PEOE_VSA4	2.3	3.8
PEOP_RPC+	0.2	0.3			
SMR_VSA5	166.0	127.9			

chemical scaffold to make the SMR\_VSA5 value falling into either the inhibitor or the noninhibitor range.

QSAR models for CYP1A2 inhibition were recently reported by Chohan et al.<sup>14</sup> using four statistical approaches. The training set consisted of 109 compounds, 22 of them being in-house compounds. A positive point of their approach was to remeasure the  $pIC_{50}$  for 81 compounds. The  $pIC_{50}$  of the remaining 28 drugs were taken from literature which gives a certain homogeneity to their biological data. The models that have been constructed combined literature results as well as in-house data and descriptors computed using an in-house application. To

**Table 7.** Descriptors Comparison for Four Compounds with Their Real Class and Their Predicted Class<sup>a</sup>

Molecule	Class	Prediction	Molecular descriptors					
			a_hyd	bpol	Chi1	VDistEq	PEOP_RPC+	SMR_VSA5
 Clomipramine	(+)	(+)	18.00	29.52	10.65	3.05	0.15	180.11
 Imipramine	(-)	(-)	17.00	30.20	10.25	3.03	0.17	197.75
 Nefazodone	(-)	(+)	22.00	46.20	16.13	4.09	0.19	196.52
 Hydroxynefazodone	(-)	(-)	22.00	46.20	16.51	4.09	0.17	177.65

<sup>a</sup> (+) stands for the inhibitors and (-) for the noninhibitors when considering a 10  $\mu$ M class threshold.

**Table 8.** Summary of the Best Predictive Models Obtained for CYP2D6 and CYP1A2 Inhibition<sup>a</sup>

CYP	dataset	threshold, $\mu\text{M}$	descriptors	Accuracy, %	Sensitivity, %	Specificity, %	Precision, %
2D6	global	10	2D+PVSA	78	82	75	70
2D6	bufuralol	3/30	2D+PVSA	88	80	98	98
2D6	dextro.	3/30	2D+PVSA	82	81	83	90
2D6	AMMC	3/30	2D+PVSA	89	82	92	82
2D6	$K_i$	10	2D+PVSA	90	88	92	90
2D6	$K_i$	3/30	PVSA	90	90	91	92
2D6	$K_i$	3/30 <sup>b</sup>	2D+PVSA	83	Precision: high = 85%, medium = 82%, poor = 83%		
2D6	$\text{IC}_{50}$	10	2D (protonation)	85	89	84	69
2D6	$\text{IC}_{50}$	3/30	2D (protonation)	85	83	89	72
2D6	global	25	3D	84	83	85	85
2D6	global	25	2D+PVSA+3D	87	88	87	87
1A2	$K_i$	30	PVSA	89	95	83	85
1A2	$K_i$	30	2D+PVSA	89	86	92	93
1A2	$\text{IC}_{50}$	30	PVSA	86	83	88	83
1A2	Global	50	3D	90	91	88	89
1A2	Global	10/200 <sup>b</sup>	2D+PVSA+3D	86	Precision: high = 83%, medium = 85%, poor = 90%		

<sup>a</sup> For each model, datasets, classes threshold, and type of descriptors are precised. <sup>b</sup> Multi-class study.

assess the diversity of the dataset, these authors used a hierarchical clustering with a database of 594 marketed drugs and then computed Euclidean distances between compounds in the training and validation test sets. The dissimilarity was evaluated based on 123 calculated descriptors, which is similar to our approach but based on chemical fingerprints. It was found that descriptors expressing lipophilicity and aromaticity were the most relevant descriptors to model CYP1A2 inhibition. The present study shows the influence of hydrophilicity expressed by logP-based descriptors present in the CYP1A2 tree as can be seen from Figure 4. In addition, size and electrostatic interactions are relevant as indicated by the SMR\_VSA6 and PEOE\_VSA+4 descriptors.

Finally, we attempted to classify some test sets used in other studies. We isolated 41 compounds from the work of Chohan et al. that were not used in our CYP1A2 sets.  $K_i$  values were calculated on the basis of the  $\text{IC}_{50}$  values and, in doing so, the prediction accuracy was 91% and 78% for inhibitors and non-inhibitors, respectively (details of the prediction are presented in the Supporting Information).

All the methods used in the already published articles present their own advantages and difficulties depending on the type of datasets and descriptors used. However, the results presented by all these authors seem to be satisfying. Nevertheless, one wishes to emphasize that RP is a simple, reliable, and validated method which is extremely easy to implement. It is the optimal tool for high-throughput screening performed at the beginning of a drug discovery process.

## Conclusions

Efficient drug developments suppose an early prediction of ADME properties. In the field of metabolism, interactions with cytochromes (CYPs) are significant. The goal of this study was to develop efficient prediction models for the inhibition of CYP2D6 and CYP1A2 using binary decision trees built with a recursive partitioning (RP) technique. Various datasets, different sets of descriptors, i.e., 2D, P\_VSA, and 3D, and diverse inhibitor/noninhibitor class thresholds were tested to obtain the best possible models. It was shown that these datasets covered a wide chemical space. To further picture chemical diversity, 2D-descriptor-based principal components analysis was performed where both training and test datasets were projected on the AurSCOPE ADME/DDI database or Specs database,<sup>32</sup> showing that the available molecules are spread over the chemical space (cf. Supporting Information). Specs database was reported to include various and diverse scaffolds and fewer duplicates compared to other commercial libraries.<sup>33</sup>

A synthetic view of the best models obtained for CYP2D6 and CYP1A2 inhibition is shown in Table 8. Exploring both CYPs in the same way, more high-quality models, with an overall Accuracy of at least 80%, were obtained with CYP2D6 datasets (11 models) in comparison to CYP1A2 (5 models). Differentiating  $K_i$  and  $\text{IC}_{50}$  measures led to the best models for both cytochromes P450, especially  $K_i$ -based models that reached an overall Accuracy of 90%. The abundance of high-structured data for CYP2D6 allowed us to build different models based on the probe substrates used in the inhibition experimental protocol. This study resulted in three specific models, with a 82% minimum Accuracy, for Bufuralol, dextromethorphan, and AMMC as probe substrates.

The use of P\_VSA descriptors was particularly efficient, and models reaching 95% of correct inhibitors classification could be generated. 3D descriptors also provided promising results but needed longer computation time, including a conformational optimization part of all the molecules. Therefore, that kind of descriptor cannot be applied easily when a high-throughput screening is needed. Two multiclass models were also generated with success despite the intrinsic difficulties of these complex approaches. Accuracy values of 83 and 86% for CYP2D6 and CYP1A2, respectively, were reached.

This work was focused on using various biological data to constitute datasets that will lead to efficient models. With the proposed models, we predicted with good performance the CYP2D6 and CYP1A2 inhibition potencies for a large series of molecules. The success of our strategy is based on the unusual quality of our data that is a main difference with other studies. Indeed, our datasets presented a good chemical diversity while also being highly structured. A deepened access to biological protocols for each measure allowed us to constitute very relevant datasets for each case. This was made possible by a complete analysis of the data coming from literature and the high structured databases available at Aureus Pharma.

The validation with external test sets led to fulfilling results. The main advantage of our RP-based method is that it is easy and quick to implement. This study also permitted us to validate the conditions of application of selected datasets. Further steps will be to use these datasets with other methods and/or descriptors.

## Materials and Methods

**Data Collection.** All the data collected for the study comes from the Aureus Pharma<sup>34</sup> knowledge databases. These databases have been designed to give access to detailed biological protocols as well as chemical data. The knowledge bases cover several domains



of pharmacology and contain large amount of measures about different systems, i.e., CYP, G-protein coupled receptors, ion channels, ... The AurSCOPE ADME/DDI knowledge database was queried to retrieve inhibition measures for CYP2D6 and CYP1A2; all the measures were extracted from 322 publications among more than 4900 references recorded in the AurSCOPE ADME/DDI database.

When several data were reported in the literature for the same compound, specific attention was given in selecting the most coherent data point, based on consistency and homogeneity of the biological protocol (biological material, substrate probe, etc.). Only values corresponding to inhibition experiments done with common probe substrates for the two CYPs were considered. Eventual duplicates were eliminated on the basis of chemical fingerprints. Following this process, the initial datasets included 498 inhibition measures for human CYP2D6 and 306 for CYP1A2, IC<sub>50</sub>, and K<sub>i</sub> values considered together.

We demonstrated significant differences with models built with K<sub>i</sub> or IC<sub>50</sub>. The K<sub>i</sub> value is an inhibition constant independent of the type or concentration of substrate and incubation conditions that define the affinity of the inhibitor for the enzyme, whereas IC<sub>50</sub> is the concentration of inhibitor required to cause 50% inhibition under a given set of experimental conditions. Using K<sub>i</sub> is always preferable rather than IC<sub>50</sub> values or percentage of inhibition.

Detailed information concerning the construction of the datasets regarding data standardization and how the quantitative biological data points (IC<sub>50</sub> vs K<sub>i</sub>, problem of multiple and sometimes discordant activities available) were selected is given in the Supporting Information.

**Descriptors.** To build the decision trees, three types of descriptors were mainly used. First, 114 two-dimensional (2D) descriptors were calculated for all the compounds related to CYP2D6 and CYP1A2 inhibition measures. The calculation of the descriptors, as implemented in the MOE software,<sup>35</sup> was based on the connectivity table of each molecule (nature of the atoms, nature of the bonds, connectivity) and on tabulated parameters. The set of 2D descriptors also contained intuitive information about the molecules such as molecular weight, number of a given atom, number of H-bond acceptors/donors, lipophilicity, etc., these descriptors supposedly leading to very interpretative decision trees.

A second type of descriptor was created with 32 P\_VSA parameters.<sup>36</sup> They are based on the approximation at atomic level of the molecular van der Waals surface area, VSA<sub>i</sub>, along with several other molecular properties, P<sub>i</sub>. VSA<sub>i</sub> values were calculated using parameters from the MMFF94 force field,<sup>37</sup> and the P<sub>i</sub> considered were the molar refractivity, logP(o/w), and the electrostatic properties or pharmacophore characteristics. Each descriptor in the series was defined to be the sum of the VSA<sub>i</sub> over all atoms *i* for which its P<sub>i</sub> value is in a specified range [*a*,*b*]. The ranges were determined by percentile subdivision over a database of 44795 compounds from the Maybridge catalog.<sup>38</sup> The descriptors were verified to be uncorrelated, and it was shown that a small numbers of those contained much more information encoded than larger sets of other popular descriptors.<sup>36</sup> It should be noted that the calculation of the P\_VSA descriptors needs only the 2D molecular connectivity as input by MOE.

The third type of descriptor was composed of 3D parameters. Considering that structures provided by literature are planar projections, an energy minimization was performed on all the molecules to generate reliable 3D coordinates. This has been completed with the CORINA software.<sup>39–42</sup> Once generated, the 3D coordinates were used by MOE to calculate the 3D descriptors.

**Computational Method.** Managing all data, calculating the descriptors, constituting the training and test sets, building the trees, and predicting the classes were all executed with MOE on a Windows computer (3 GHz CPU, 1 GB RAM). The MOE QuaSAR—Classify module that implements RP algorithm was used to build and display the classification trees. For choosing the output tree, a cross-validation approach was used; the *k*-parameter was 2, meaning that the algorithm subdivides the whole dataset in two

equal parts. A training set was randomly chosen in one set of data and used to perform the tree building. The rest of the data is then used as test set to confirm the relevance of the tree, both sets remaining mutually exclusive. To avoid the overtraining during the tree growing, QuaSAR Classify uses a pruning process. A sequence of subtrees is constructed from the initial tree, and the test dataset is used to choose the final output tree from this sequence. Pruning removes one or more branches of a tree. The roots of the branches to be removed remain part of the pruned tree, becoming leaf nodes.

MOE default parameters were used to build the RP trees. The node split was set to 10, meaning that, once a branch of a tree contains 10 or less compounds, it cannot be further subdivided. Thus, each branch becomes a terminal leaf to which a class is attributed, i.e., either inhibitor or noninhibitor. The maximum depth of the trees was set to 10, but it was not a real restriction, as the maximum depth observed was 7.

To compare the performance of the different trees, several measures were used. Accuracy (eq 1) is the overall classification accuracy of a prediction model; it corresponds to the ratio of correctly classified compounds.<sup>43</sup> Misclassification rate, known as *R(t)*, associated directly by MOE to each of the built tree, represents the ratio of incorrectly classified compounds. It obviously means that the Accuracy equals 1 - *R(t)*. Sensitivity (also known as Recall) (eq 2) is the ratio of inhibitors correctly predicted, whereas Specificity (eq 3) is the ratio of noninhibitors correctly predicted. Precision (eq 4) is a measure of the ability of a tree to predict a specific class. In this study, only the Precision of the inhibitor class was considered.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

where TP = number of true positives, TN = number of true negatives, FP = number of false positives, and FN = number of false negatives.

As the goal of this study was the construction of models for the prediction of the inhibition of CYP2D6 and CYP1A2, we paid particular attention to optimize the Sensitivity of the obtained decision trees.

**Acknowledgment.** J.B. wishes to thank the “Fond pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (FRIA) for his Ph.D. fellowship. All authors acknowledge the members of the Aureus Pharma’s Knowledge Management ADME/DDI team for searching, analyzing, and recording the data. We also thank all the people involved in the management and the extraction of the data. The CORINA demo version was gracefully provided by Molecular Networks.<sup>42</sup>

**Supporting Information Available:** Description of the diversity of our training and test sets compared to the Specs database; a precise description of how our datasets were selected (data sets identification, construction, dealing with K<sub>i</sub> and IC<sub>50</sub>, standardization); details for the prediction of the test sets related to other studies; correlation matrices of the descriptors involved in the best models; the list of the descriptors used in this study as available from MOE, which is also accessible; the training data sets used to build the best models, i.e., 163 and 81 K<sub>i</sub> measures for CYP2D6 and CYP1A2, respectively. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nature Rev. Drug Discovery* 2004, 3, 711–715.

- (2) Nelson, D. R.; Koymans, L.; Kamataki, T.; Stegeman, J. J.; Feyereisen, R.; Waxman, D. J.; Waterman, M. R.; Gotoh, O.; Coon, M. J.; Estabrook, R. W.; Gunsalus, I. C.; Nebert, D. W. P450 Superfamily: Update on New Sequences, Gene Mapping, Accession Numbers and Nomenclature. *Pharmacogenetics* **1996**, *6*, 1–42.
- (3) Tredger, J. M.; Stoll, S. Cytochromes P450—Their Impact on Drug Treatment. *Hosp. Pharmacist* **2002**, *9*, 167–173.
- (4) Farrel, G. C. *Drug-Induced Liver Disease*, 1st ed.; Churchill Livingstone: New York, 1994.
- (5) Michalets, E. L. Clinically Significant Cytochrome P-450 Drug Interactions. *Pharmacotherapy* **1998**, *18*, 84–112.
- (6) Ekins, S.; Wrighton, S. A. Application of *in Silico* Approaches to Predicting Drug-Drug-Interactions. *J. Pharm. Toxicol. Methods* **2001**, *45*, 65–69.
- (7) van de Waterbeemd, H.; Gifford, E. ADMET *in Silico* Modelling: Towards Prediction Paradise? *Nature Rev. Drug Discovery* **2003**, *2*, 192–204.
- (8) 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.
- (9) Korolev, D.; Balakin, K. V.; Nikolsky, Y.; Kirillov, E.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Nikolskaya, T. Modeling of Human Cytochrome P450-Mediated Drug Metabolism Using Unsupervised Machine Learning Approach. *J. Med. Chem.* **2003**, *46*, 3631–3643.
- (10) Haji-Momenian, S.; Rieger, J. M.; Macdonald, T. L.; Brown, M. L. Comparative Molecular Field Analysis and QSAR on Substrates Binding to Cytochrome P450 2D6. *Bioorg. Med. Chem.* **2003**, *11*, 5545–5554.
- (11) Kemp, C. A.; Flanagan, J. U.; van Eldik, A. J.; Maréchal, J.-D.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. F. Validation of Model of Cytochrome P450 2D6: An *in Silico* Tool for Predicting Metabolism and Inhibition. *J. Med. Chem.* **2004**, *47*, 5340–5346.
- (12) Korhonen, L. E.; Rahnasto, M.; Mähönen, N. J.; Wittekindt, C.; Poso, A.; Juvonen, R. O.; Raunio, H. Predictive Three-Dimensional Quantitative Structure–Activity Relationship of Cytochrome P450 1A2 Inhibitors. *J. Med. Chem.* **2005**, *48*, 3808–3815.
- (13) 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.
- (14) 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.
- (15) 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.
- (16) Koziol, J. A.; Zhang, J. Y.; Casiano, C. A.; Peng, X. X.; Shi, F. D.; Feng, A. C.; Chan, E. K. L.; Tan, E. M. Recursive Partitioning as an Approach to Selection of Immune Markers for Tumor Diagnosis. *Clin. Cancer Res.* **2003**, *9*, 5120–5126.
- (17) Rusinko, A., III; Farnen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017–1026.
- (18) Jones-Hertzog, D. K.; Mukhopadhyay, P.; Keefer, C. E.; Young, S. S. Use of Recursive Partitioning in the Sequential Screening of G-Protein-Coupled Receptors. *J. Pharm. Toxicol.* **1999**, *42*, 207–215.
- (19) Godden, J. W.; Furr, J. R.; Bajorath, J. Recursive Median Partitioning for Virtual Screening of Large Databases. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182–188.
- (20) van Rhee, A. M. Use of Recursion Forests in the Sequential Screening Process: Consensus Selection by Multiple Recursion Trees. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 941–948.
- (21) Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276–285.
- (22) Jchem, version 3.0.10, ChemAxon, 1037 Budapest, Hungary, www.chemaxon.com.
- (23) 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.
- (24) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A Novel Approach to Predicting P450 Mediated Drug Metabolism. CYP2D6 Catalyzed N-Dealkylation Reactions and Qualitative Metabolite Predictions Using a Combined Protein and Pharmacophore Model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (25) de Graaf, C.; Vermeulen, N. P. E.; Feenstra, K. A. Cytochrome P450 in Silico: An Integrative Modeling Approach. *J. Med. Chem.* **2005**, *48*, 2725–2755.
- (26) Kalgutkar, A. S.; Zhou, S.; Fahmi, O. A.; Taylor, T. J. Influence of Lipophilicity on the Interactions of *N*-Alkyl-4-phenyl-1,2,3,6-tetrahydropyridines and Their Positively Charged *N*-Alkyl-4-phenylpyridinium Metabolites with Cytochrome P450 2D6. *Drug Metab. Dispos.* **2003**, *31*, 596–605.
- (27) Vaz, R. J.; Nayeem, A.; Santone, K.; Chandrasena, G.; Gavai, A. V. A 3D-QSAR Model for CYP2D6 Inhibition in the Aryloxypropanolamine Series. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3816–3820.
- (28) 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.
- (29) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indices and Kappa Shape Indices in Structure–Property Modeling. *Revi. Comput. Chem.* **1991**, *2*, 367–422.
- (30) Hall, L. H.; Kier, L. B. The Nature of Structure–Activity Relationships and Their Relation to Molecular Connectivity. *Eur. J. Med. Chem.—Chim. Ther.* **1997**, *4*, 307–312.
- (31) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmacophore and Three-Dimensional Quantitative Structure Activity Relationship Methods for Modeling Cytochrome P450 Active Sites. *Drug Metab. Dispos.* **2001**, *29*, 936–944.
- (32) Specs, Delft, Holland, <http://www.specs.net>.
- (33) Baurin, N.; Baker, R.; Richardson, C.; Chen, I.; Foloppe, N.; Potter, A.; Jordan, A.; Roughley, S.; Parratt, M.; Greaney, P.; Morley, D.; Hubbard, R. E. Drug-like Annotation and Duplicate Analysis of a 23-Supplier Chemical Database Totaling 2.7 Million Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643–651.
- (34) Aureus Pharma, Paris, 75010, France, <http://www.aureus-pharma.com/>.
- (35) MOE, Chemical Computing Group Inc., Montreal, H3A 2R7 Canada, <http://www.chemcomp.com>.
- (36) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.
- (37) Halgren, T. A *Merck Molecular Force Field: I. Basis, Form, Scope, Parameterization, and Performance of MMFF94*, *J. Comput. Chem.* **1996**, *17*, 490–519.
- (38) Maybridge Chemical Company Ltd., Cronwall, PL34 0HW England, <http://www.maybridge.com>.
- (39) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D Coordinates for Organic Molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537–547.
- (40) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
- (41) Sadowski, J.; Schwab, C. H.; Gasteiger, J. 3D Structure Generation and Conformational Searching. In *Computational Medicinal Chemistry and Drug Discovery*; P. Bultinck, H. De Winter, W. Lange-naecker, J. P. Tollenaere, Ed.; Dekker Inc.: New York, 2003, pp 151–212.
- (42) Molecular Networks GmbH, 91052 Erlangen Germany <http://www2.chemie.uni-erlangen.de/software/corina/index.html>; <http://www.mol-net.com/>.
- (43) Jacobsson, M.; Lidén, P.; Stjernschantz, E.; Boström, H.; Norinder, U. Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* **2003**, *46*, 5781–5789.

JM060267U