Virtual Screening for Cytochromes P450: Successes of Machine Learning Filters

Julien Burton[*,1], Ismail Ijjaali[2], François Petitet[2], André Michel[2] and Daniel P. Vercauteren[1]

[1] *Laboratoire de Physico-Chimie Informatique, Groupe de Chimie Physique, Théorique et Structurale, University of Namur (FUNDP), 61 rue de Bruxelles, B-5000 Namur, Belgium*

[2] *AUREUS-PHARMA, 174 quai de Jemmapes, F-75010 Paris, France*

Abstract: Cytochromes P450 (CYPs) are crucial targets when predicting the ADME properties (absorption, distribution, metabolism, and excretion) of drugs in development. Particularly, CYPs mediated drug-drug interactions are responsible for major failures in the drug design process. Accurate and robust screening filters are thus needed to predict interactions of potent compounds with CYPs as early as possible in the process. In recent years, more and more 3D structures of various CYP isoforms have been solved, opening the gate of accurate structure-based studies of interactions. Nevertheless, the ligand-based approach still remains popular. This success can be explained by the growing number of available data and the satisfying performances of existing machine learning (ML) methods. The aim of this contribution is to give an overview of the recent achievements in ML applications to CYP datasets. Particularly, popular methods such as support vector machine, decision trees, artificial neural networks, *k*-nearest neighbors, and partial least squares will be compared as well as the quality of the datasets and the descriptors used. Consensus of different methods will also be discussed. Often reaching 90% of accuracy, the models will be analyzed to highlight the key descriptors permitting the good prediction of CYPs binding.

Keywords: Cytochrome P450, ADME, *in silico* filters, machine learning.

INTRODUCTION

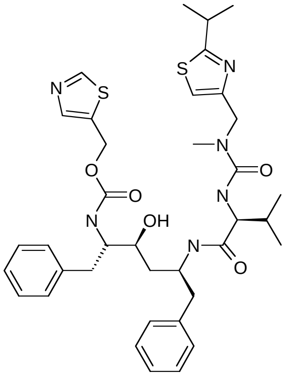
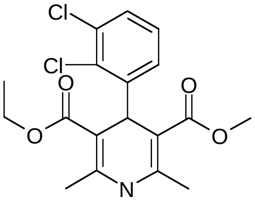
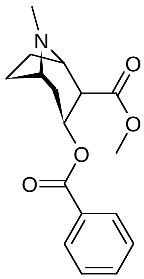
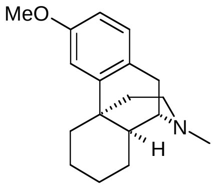
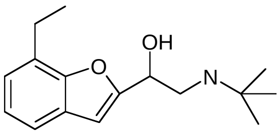
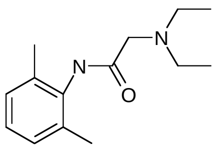
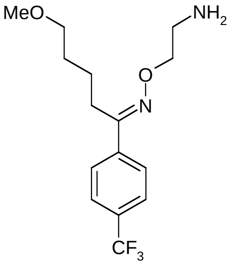
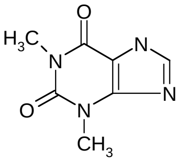
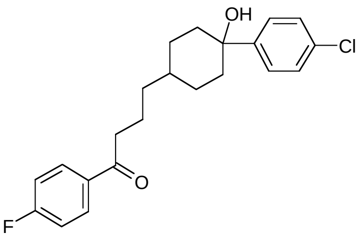
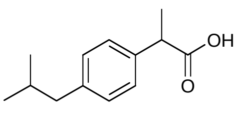
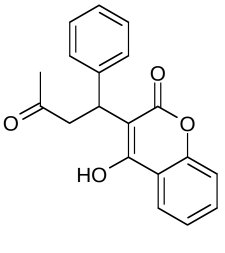
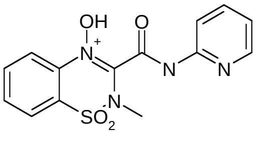
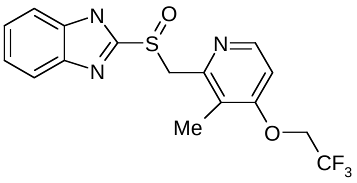
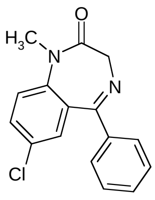
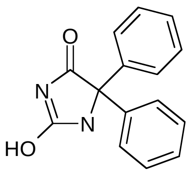
Cytochromes P450 (CYPs) are a superfamily of enzymes containing a heme prosthetic group and a polypeptide. The superfamily is organized and divided in families with sequence homology > 40 % [1]. The heme moiety is the oxidation reaction center and the apoprotein being involved in substrate selectivity and specificity. Large polymorphism exists which explains the individual susceptibility and population differences in metabolism. CYPs are located throughout the body but are predominant in location where exposition to external small molecules, *i.e.*, xenobiotics such as drugs, environmental substances, alimentary component, occurs. CYPs are particularly abundant in liver, intestine, lung, and kidneys. The overall goal of the catalyzed enzymatic reaction is to favor the elimination of hydrophobic compounds by adding polar groups and by this way self protecting the body against external compounds that are eliminated faster. Oxidation reactions are held with NADPH or NADH as co-enzyme. In the reaction, the role of xenobiotics can be diverse. They could act as substrate, inducer or inhibitor. Inhibition could be competitive (reversible) or irreversible due to the formation of an active metabolite which tightly binds to the active site generating a long lasting inhibition [2]. Tables **1** and **2** contain examples of substrates and inhibitors for the major CYP isoforms. For their importance in metabolism, studies of interactions with CYPs are significant in the ADME (absorption, distribution, metabolism, excretion) stage of drug discovery programs. Seven CYPs are involved in the metabolism of more than 90 % of the current drugs in use and in clinical trials: 3A4, 2D6, 1A2, 2C9, 2C19, 2E1, and 2C18. The first five are the most important and several show polymorphism [3].

CYPs have a strong impact on pharmacokinetic and drug-drug interactions (DDI). They have influences on bioavailability, clearance, elimination, and almost all pharmacokinetic parameters. They also have an impact on toxicity, for example, when a reactive metabolite is created. Moreover, CYPs mediated metabolism is the major substratum for DDI, *i.e.*, one drug inhibiting the metabolism of the other and therefore increasing the plasma concentration [4]. Lots of strategies are in place in drug discovery to reduce attrition due to poor ADME compound characteristics [5]. Important progresses have been made in the miniaturization and possible throughput of *in vitro* tests. Early assessment of ADME properties allows selecting for preclinical “drugable” compounds [6]. Inhibition and induction capabilities as well as the metabolic stability of each compound are now particularly studied and testing compound interactions with the major CYP isoforms is now systematic in the drug development process. In addition, to avoid deleterious effect on CYPs inhibition or induction the early assessment of potential DDI can be done on the basis of *in vitro* data. But the most efficient strategy in terms of speed and cost is the application of *in silico* filters to screen large databases of compounds. X-ray structures become available and encourage *in silico* structure-based approach [7-9]. Nevertheless, ligand-based methods remain a powerful tool in the filtering of databases. As data increase from year to year, it is easier to build reliable models to predict CYPs binding [10]. In this field, machine learning (ML) has been applied in numerous studies that occupy a large place in literature.

[*] Address correspondence to this author at the Laboratoire de Physico-Chimie Informatique, Groupe de Chimie Physique, Théorique et Structurale, University of Namur (FUNDP), 61 rue de Bruxelles, B-5000 Namur, Belgium; Tel: +32 (0)81 72 54 62; Fax: +32 (0)81 72 54 66;
E-mail: julien.burton@gmail.com

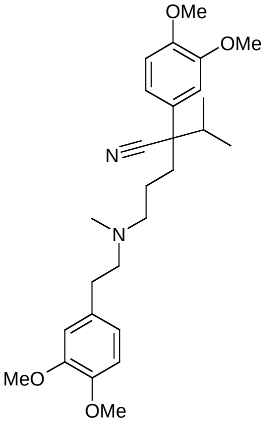
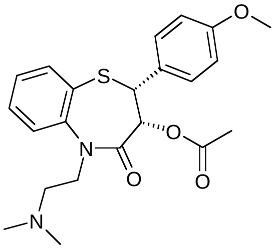
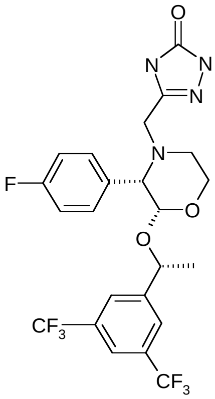
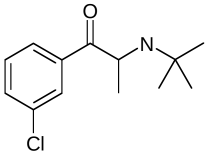
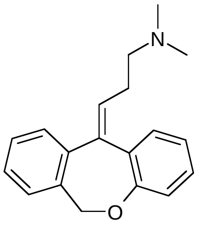
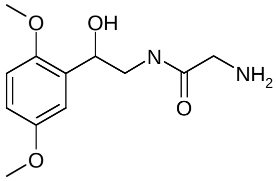
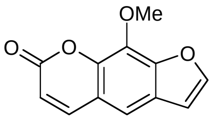
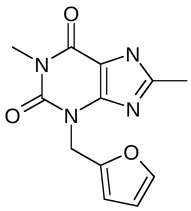
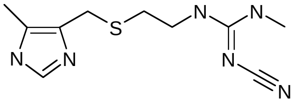
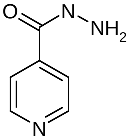
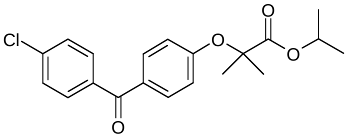
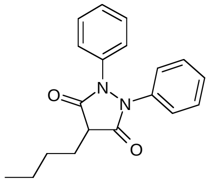
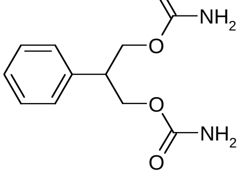
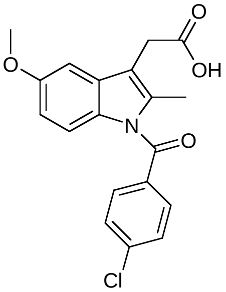
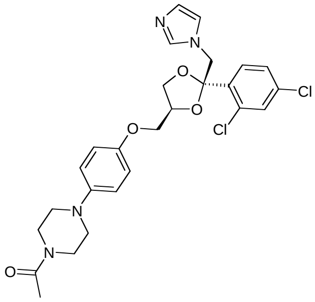
1386-2073/09 $55.00+.00

Table 1. Structural Formula of Several Typical Drugs Reported as Substrates of the 5 Main CYP Isoforms, CYP3A4, CYP2D6, CYP1A2, CYP2C9, and CYP2C19. For a Larger Collection of Compounds, Refer to [122]

CYP Isoform	Substrates
CYP3A4 (~36 %)[1]	 Ritonavir  Felodipine  Cocaine
CYP2D6 (~19 %)[1]	 Dextromethorphan  Bufuralol  Lidocaine
CYP1A2 (~11 %)[1]	 Fluvoxamine  Theophylline  Haloperidol
CYP2C9 (~10 %)[1]	 Ibuprofen  Warfarin  Piroxicam
CYP2C19 (~8 %)[1]	 Lansoprazole  Diazepam  Phenytoin

Proportion of drug metabolized by the CYP in the human liver; data from reference [1].

Table 2. Structural Formula of Several Typical Drugs Reported as Inhibitors of the 5 Main CYP Isoforms, CYP3A4, CYP2D6, CYP1A2, CYP2C9, and CYP2C19

CYP Isoform	Inhibitors
CYP3A4	 Verapamil  Diltiazem  Aprepitant
CYP2D6	 Bupropion  Doxepin  Midodrine
CYP1A2	 Methoxsalen  Furafylline  Cimetidine
CYP2C9	 Isoniazid  Fenofibrate  Phenylbutazone
CYP2C19	 Felbamate  Indomethacin  Ketoconazole

For a larger collection of compounds, refer to [122].

This review's target is to show how CYPs inhibitors and substrates are modeled by different methods. The first part of the work browses the common ML methods stressing on inherent particularities, problems, and improvements brought by the authors of the discussed studies. Comparisons and generalities common to all methods will be discussed in the Discussion part, such as the datasets, the descriptors choice, and the controlling factors affecting CYPs interactions. All the studies reported herein are summarized in Table **3** for the inhibition data and Table **4** for the substrates. Table **5** comprises the definition of the performance parameters used by the authors.

MAIN RESULTS FOR DIFFERENT METHODS

Support Vector Machine (SVM). SVM was initially introduced in 1992 by Vladimir Vapnik [11] and appeared in the chemistry domain in the early 2000 [12]. Over the last decade, this method has been gaining considerable attention in the ML community since it has been applied successfully in several real-world applications. In pharmaceutical research, the applications included predicting activity toward therapeutic targets, ADME effects, or avoiding adverse drug reactions [12-17]. For an extensive description of theory, one can refer to [18].

The general principle of SVM is to perform a classification by constructing an *n*-dimensional hyperplane that optimally separates the data into two categories. Non-linear problems are treated with the help of kernel functions. The main types of kernel functions encountered are polynomial, radial basis functions, and sigmoid (tanh). SVM presents several advantages. It minimizes the empirical classification error and maximizes the geometric margin; therefore, it is also known as a maximum margin classifier. Moreover, SVM is based on the structural risk minimization principle which allows the building of predictive models, even if the descriptors are numerous and redundant. These classifiers are, thus, very accurate, even for high dimensionality problems. Let us mention that SVM theory has also been applied to regression problems (support vector regression) [19].

In 2005, Kriegl *et al.* [20] proposed a way to select the optimum SVM parameters, the width γ of radial basis functions and the C penalty parameter, in the context of CYP3A4 inhibition. The parameter C determines how to penalize missclassifications. The final model is actually a tradeoff between a large margin and a small error of classification. The authors browsed the (C, γ) space to determine the optimal parameter pairs. The analysis of the 3D space defined by C, γ, and the average 10-fold cross validation prediction (Q^2) permitted to define a triangular-shaped region of optimal SVM parameters. That region became smaller as the classification problem was more complex (3 classes). The training set was composed of 807 compounds described by the MOE descriptors [21]. The best model was obtained with a Gaussian kernel and had an accuracy of 94% for an external test set. A 3-classes model reached 75% of accuracy. A blind test was performed on 7 drugs and only one was predicted as strong inhibitor whereas it was actually a weak inhibitor. The systematic exploration of the (C, γ) space is thus a valuable strategy to exploit all the power of SVM and build very predictive models.

Very recently, Eitrich *et al.* [22] addressed the problem of unbalanced datasets and applied their method to SVM and maximum entropy modeling. The training set contained 185 CYP2D6 inhibitors described by 557 various descriptors and 458 bits E_screen strings [23]. Both Gaussian and Slater kernels were used with similar results. The authors considered threshold moving and oversampling to raise the accuracy of the models based on highly unbalanced data. To preprocess the data, the oversampling technique increases the number of positive examples, by a factor 3 in this particular case, and does not lead to any loss of information. The post-processing is handled by the threshold moving method in which the output threshold is moved toward the inexpensive (major) class. Threshold moving permitted to influence positively the sensitivity of the models, while oversampling improved the overall results. The best result, *i.e.*, a sensitivity of 92%, was obtained with the combination of both techniques. Consequently, one can conclude that both oversampling and threshold moving successfully treated the problem of unbalanced datasets which is often encountered in CYPs datasets.

Yap and Chen [24] used two SVM consensus methods: positive majority consensus SVM (PM-CSVM) and positive probability CSVM (PP-CSVM) to study inhibitors and substrates of CYP3A4, CYP2D6, and CYP2C9. 702 compounds were defined as inhibitors or non-inhibitors and substrate or non-substrates for each of the three CYPs. Nearly all the models presented Matthews correlation MCC [25] above 0.800 for the external validation with a maximum of 0.899 for the CYP3A4 substrates. It appeared that consensus methods were better with an MCC value of approximately 0.1 units above single SVM classifiers. The authors also classified the datasets with other methods; SVM was highly superior to the other methods such as multiple linear regression (MLR) (MCC = 0.586), logistic regression (0.555), PLS (0.528), decision tree (0.423), and *k*-nearest neighbors (0.759).

Merkwirth *et al.* [26] compared several methods, *i.e.*, single SVM, an ensemble of 15 SVM, ensemble of and *k*-nearest neighbors (*k*-NN) classifiers, and ridge regression with a strong emphasis on SVM. Exploring different combinations of methods and 1814 structural descriptors, they concluded that the ensemble of 15 SVM classifiers with a restricted set of descriptors was the best compromise between results and computational complexity. Nevertheless, the best model, in terms of performance, was the single SVM classifier; but it is said to be more demanding in terms of computation resources as it was based on the entire pool of descriptors. It reached, on the training set, a MCC value of 0.87, *versus* 0.86 for ensemble SVM. They also obtained a MCC value of 0.62 for the single SVM classifier with an external validation set. Another remarkable point of their study was the selection of optimal SVM parameters (γ and C) by the out-of-train technique (OOT), which can be considered as an enhanced version of traditional cross validation.

Decision Trees (DT). DTs, also commonly known as recursive partitioning (RP), are usually used to design interpretable and rapid filters [27,28]. The typical structure of a DT consists of a root node linked to two or more child

Table 3. Summary of the SVM, DT, ANN, and *k*-NN CYPs Inhibition Studies Discussed in the Review

Method	CYP	Measure	Training Set	Test Set	Type of Descriptors	Performance (Test)
Support Vector Machine (SVM)						
SVM [20]	3A4	IC_{50}	807	538	2D, QM, surface, in-house	$C^2 = 0.75$ Accuracy = 94%
SVM for unbalanced datasets [22]	2D6	IC_{50}	185	78	E_screen + various	Sensitivity = 92%
PM-CSVM PP-CSVM [24]	3A4 2D6 2C9	K_i	702	100	Structural and chemical	MCC =0.893 0.821 0.835
SVM + OOT technique [26]	3A4	IC_{50}	410	85	2D, Ghose and Crippen, topological, electronic	MCC =0.62
SVM [84]	3A4	K_i	4000	470	MACCS keys	κ = 0.62
Decision Trees (DT)						
RP [34]	3A4 2D6	% inh.	1759	98	Augmented atoms	Spearman's *ρ* = 0.61 0.49
Ensemble RP [29]	2D6	K_i	100	51	Various 2D	Accuracy = 80%
Line-Walking RP [37]	3A4 2D6 2C9	K_i	702	100	Shape and surface charge	Accuracy = 85.0% 86.6% 90.6%
RP [38]	1A2 2D6	K_i and IC_{50}	498 306	58 34	2D and 3D (MOE)	Accuracy = 81% 89%
CART [80]	1A2	pIC_{50}	109	68	Various	RMSE = 1.0
RP [84]	3A4	K_i	4000	470	MolconnZ	κ = 0.62
Artificial Neural Networks (ANN)						
ANN [52]	3A4	IC_{50}	218	72	2D Unity fingerprints	Sensitivity = 91.7%
ANN [53]	2D6	K_i	1810	600	E-state and Barnard fingerprints	κ = 0.58
Bayesian-ANN [80]	1A2	pIC_{50}	109	68	Various	RMSE = 1.3
k-Nearest Neighbors (k-NN)						
Gaussian kernel weighted *k*-NN [63]	2D6 3A4	% inh.	865 1037	288 345	FCFP and ECFP	Accuracy = 82% 87%
k-NN [84]	3A4	K_i	4000	470	MolconnZ	κ = 0.58
Partial Least Squares (PLS)						
PLS [69]	3A4	K_i	11	/	MS-WHIM (surface, size, shape)	$q^2 = 0.32$ (3D) [1] 0.44 (4D) [1]
PLS [72]	3A4	IC_{50}	53	9	2D topological	$r_{pred} = 0.744$ $s = 0.769$
PLS - Discriminant Analysis [73]	3A4	IC_{50}	967	/	2D, QM, surface, in-house	Accuracy = 66.0%
PLS [74]	2D6	K_i	64	50	Grid-independent	Accuracy = 78%
PLS [75]	1A2	IC_{50}	46	/	CoMFA	$r^2 = 0.87$ (CoMFA) [1] 0.90 (GRID) [1] (regression for the training set)
PLS and MLR [80]	1A2	pIC_{50}	109	68	Various	RMSE = 1.3 (PLS) 1.4 (MLR)
Consensus of Different Methods						
ANN + Bayesian [53]	2D6	K_i	1810	600	E-state and Barnard fingerprints	κ = 0.97
PLS + MLR + Bayesian ANN + CART [80]	1A2	pIC_{50}	109	68	Various	RMSE = 1.2
RP + SVM [84]	3A4	K_i	4000	470	MACCS keys, MolconnZ, Barnard chemical information	κ = 0.83

[1] Results for the regression of the training set.

Only the best model for each method or CYP in each paper is presented. Performances are calculated on external validation sets (except if mentioned).

Table 4. Summary of CYPs Substrate Prediction Studies Discussed in the Review

Method	CYP	Training Set	Test Set	Type of Descriptors	Performance (Test)
SVM [24]	3A4 2D6 2C9	702	100	Structural and chemical	MCC =0.899 0.884 0.872
RP [36]	*in vitro* intrinsic clearance (Cl_u)	875	41	/	Spearman's ρ = -0.64
SOM [58]	38 CYPs	605	202	Various 2D	Acc.=59.9% (substrates) 64.8% (products)
SOM [59]	12 CYPs (K_m)	491	15	Various 2D	Accuracy = 87% (non-substrates)
Bayesian-regularized ANN [60]	3A4 (K_m)	44	15	E-state	Standard prediction error = 0.15 ± 0.02
PLS [76]	2D6 (K_m)	24	15	CoMFA	r^2 = 0.62 (regression for the training set)

Only the best model for a single method or CYP in each paper is presented. Performances are calculated on external validation sets.

nodes, themselves linked to other grandchild nodes, etc… The tree's unlinked nodes are called leaves and determine the classification on the basis of the major class present in it. The path of prediction can be seen as a succession of "*if…then*" decisions leading to a leaf which attribution will assign the class to the predicted compound. During the construction, RP algorithms find the more discriminating descriptor that would split the training set into two purer sets. The calculation is repeated for the purer sets to split them in even purer subsets, until termination conditions are reached, *i.e*., generally a minimal number of compounds in terminal leaves or a maximum depth of the tree. RP is known to be sensitive to the descriptors used, to unbalanced training sets, and to the composition of datasets [29]. Though, DTs are still very interpretable as they give access to critic descriptors and thresholds that would govern the classification. Moreover, the method requires less computational power and time (compared to SVM for example); hence, their use in various classification problems [30-32]. We also here note that the random forest technique becomes more and more popular and consists in a combination of single tree classifiers [33]. The general approach to derive predictions from few simple "*if…then*" conditions can be applied to regression problems. Then, the terminal nodes are not assigned to a class but to a particular value of the continuous parameter to predict.

Ekins *et al*. [34] presented a typical study for designing rapid and simple filters for CYP2D6 and CYP3A4 inhibitors. Using atom augmented descriptors and models of 20 random trees, they could reach interesting results for the training set, *i.e*., r^2 = 0.88 for CYP2D6 and 0.82 for CYP3A4 and statistically significant Spearman's ρ rank (describing the rank order and not the inhibition value itself) of 0.61 for CYP2D6 and 0.49 for CYP3A4. Based exclusively on commercially available data (~ 1750 compounds) and software (Chemtree [35]), that efficient strategy is more suited for industries. The ranking approach can be interesting to identify compounds with highest potencies of development during the drug discovery process.

Ekins [36] used again RP to design a more general filter, based on several CYPs by predicting the *in vitro* intrinsic clearance (Cl_u). A dataset of 875 molecules with human metabolic stability data was modeled with a value of r^2=0.71 for the training set. A test set of 41 compounds was classified with Spearman's ρ of –0.64 and r^2 = 0.34. Such a filter is indicated to predict a significant part of the general metabolism of a compound instead of interactions with a single CYP.

Susnow *et al.* [29] developed an ensemble approach to RP trees. Predictions were made using an average of numerous trees, avoiding the problems occurring in single trees (overfitting, correlations, …). Two-dimensional descriptors were selected using an in-house algorithm. Out of the 100 compounds in the training set, 75 were correctly classified with the ensemble method, versus 62 for standard RP. An external validation confirmed these trends by correctly predicting 80% of a test set for ensemble trees and 71% for standard RP. This demonstrates well the superiority of their ensemble classifiers. Nevertheless, both models correctly predicted 100% of the strong inhibitor compounds of the test set.

In 2006, Hudelson *et al*. [37] developed a very efficient adaptation of the RP method. The proposed line-walking RP (LWRP) differs from traditional RP by incorporating elements from SVM and, thus, producing simpler trees with lower depth and leaves. All nodes of a decision tree are considered as a hyperplane splitting the dataset in purer subsets. By incorporating the SVM methodology, LWRP is able to find more efficiently the relevant hyperplanes and, thus, may produce very optimized trees. Another huge advantage is that LWRP allows to use a remarkably low number of descriptors. In their study, only 9 descriptors depicting shape and surface charge were selected. Using the datasets of Yap and Chen [24] (see the SVM section), they obtained accurate models with accuracies of 90.6% for CYP2C9, 86.6% for CYP2D6, and 85.0% for CYP3A4. They even reached 95.6% accuracy with an improved 2C9 dataset, *i.e.* containing more accurate inhibition measures from literature and correct structures. Interestingly, they established by PCA that the nearest neighbors were not dominant in this particular classification problem. LWRP is therefore a promising evolution of RP.

The strategy of Burton *et al*. [38] was based on the particular attention paid to the quality of the data instead of any adaptation of the method. The data extracted from literature by Aureus-PHARMA [39] were diversified and activity classes were carefully assigned as lots of compounds were given several inhibition data. Exploring different datasets, RP parameters, descriptor pools and activity thresholds, the authors obtained numerous good models with accuracy > 80% for the training set. Models based on K_i values were better than those built on IC_{50} measures. The best tree for each CYP validation set corresponded to 88% of accuracy for CYP2D6 and 81% for CYP1A2. The results prove that using a very simple standard method can be successful when the data is of superior quality compared to traditional collected data. Optimum results cannot be achieved by improving the methods only.

In general, RP seems to become more and more efficient due to the improvements on the method in various domains, as LWRP, recursion forests [40], genetic algorithms (GA) [41], median partitioning [42], evolutionary programming [43], as well as on the growing quality of the data available. It is quite pleasing as that method was considered only as a fast and simple technique but sometimes a bit inaccurate compared to SVM for example.

Artificial Neural Networks (ANN). Artificial neural networks consist of mathematical models and algorithms that mimic the information processing and knowledge acquisition of the human brain. An artificial neuron collects a series of input signals and transforms them into an output signal *via* a transfer function. Basically, two classes of ANN can be distinguished: perceptrons [44-46] and Kohonen self organizing maps (SOM) [47]. A perceptron is a weighted linear combination of non-linearly transformed inputs. The output of a layer of perceptrons can be used as an input of another layer of perceptrons. Such an architecture is the basis of an ANN. Regarding SOM, artificial neurons learn to map points in an input space to coordinates in an output space. The input space can have different dimensions and topology from the output space, and the SOM will attempt to preserve those. As ANN are non-linear statistical modeling tools, they have been employed to model complex relationships between inputs and outputs. We noted a lot of applications in drug discovery [48]; these include analysis of multi-dimensional data [49], classification and prediction of biological activity and ADME properties [50], and lead discovery [51].

Molnar *et al*. [52] carried out a standard application of ANN on 290 CYP3A4 inhibitors using the 992-bits unity fingerprints. The obtained model had a 992-31-1 architecture for the input, hidden, and output layers respectively. The associated prediction ability was 97% for inhibitors and 95% for non-inhibitors from the training set (considering a borderline scoring of 0.5) and 91.7% and 88.9% for the test set. The model has then been successfully applied to the correct calculation of inhibitory indices for 8 of 9 drug candidates synthesized at Eli Lilly.

O'Brien *et al*. [53] proposed a development of ANN strategy combined to a Bayesian model. In their work, Barnard [54] and E-state [55,56] fingerprints described 2410 CYP2D6 inhibitors and non-inhibitors. The authors used a weighted κ index which is considered as a reliable index to characterize a model; it assesses the improvement of a model in prediction compared to random predictions [57]. The ANN model classified the test set of 600 compounds with κ = 0.58. They also built a Bayesian classifier based on functional class fingerprints (FCFP) and other descriptors. It reached a κ of 0.51 on the external compounds. There was a large degree of overlap between both models but several compounds were accurately defined in one but not in the other. Hence, the consensus of the models permitted to outperform the single models with a κ value of 0.60, and even 0.97 if unpredicted compounds were removed. Let us mention that, in the same article, the strategy was applied with nearly the same success to another ADME topic, namely hERG channel blocking.

Interestingly, ANN, and particularly SOM, have been shown to be one of the most popular methods to classify CYPs substrates. It can be explained by the greater diversity encountered in substrates compared to inhibitors since CYPs role is to catalyze the metabolic transformations of many diverse substances. Therefore, SOM are a powerful tool to clearly visualize the distribution of many different families of compounds.

In this topic, Korolev *et al*. [58] used Kohonen SOM to assess the probability of compounds to be transformed by CYPs. They used a substantial training set of 2200 reported substrates, non-substrates, and products for 38 different CYPs. The compounds were first described by 60 descriptors that were reduced to 7 descriptors by PCA. Unsupervised learning was used to observe the repartition of substrates, non-substrates, and products of substrates degradation and possible overlapping areas on a 10×10 map. By assigning different areas of the map to substrates or product classes, they obtained a model that classified 76.7% of substrates and 62.7% of products. Applied to a validation set, the predictions were 59.9% for substrates and 64.8% for products. That kind of prediction is certainly harder to perform than those based on inhibition because substrates and even more non-substrates and substrate degradation products cannot be predicted from structural similarities as reliably as inhibitors are.

Balakin *et al*. [59] performed a similar study with a more specific dataset of 491 K_m measures for 12 CYPs and well discriminated classes (low K_m <10μM and high K_m >100μM). Again, 60 various descriptors were reduced by PCA to 6 significant ones. They explicitly focused on the problem of overlapping inhibitors and substrates. A set of 33 CYP3A4 competitive inhibitors were projected on the substrate SOM and 31 of them (94%) were located in the low K_m area, which is what was expected. The validation was ensured by 15 other CYP3A4 inhibitors and, once more, 13 of them (87%) were classified as low K_m. One can conclude that SOM can be used with a certain success to predict inhibitors but less for substrates or products as Korolev *et al*. did.

Wang *et al*. [60] tackled the problem of K_m modeling with Bayesian-regularized ANN (BRNN). BRNN are multilayer feed-forward ANN trained with a Bayesian algorithm. Compared to classical ANN, Bayesian training produces a posterior distribution over networks weights [61]. The 59 CYP3A4 compounds were depicted by 50 E-state indices [55,56] encoding information about topological environment of atoms but also on electronic interactions from other atoms in the molecule. Tanh-sigmoid and a linear transfer function were used for the hidden and output layers; the final model had a 14-13-1 architecture. Standard error of estimation (SEE) was 0.11 ± 0.02 for the training set and standard error of prediction (SEP) was 0.15 ± 0.02 for the test set (for data scaled from 0 to 1). As SEE and SEP are in the same range of magnitude, it indicates that the model is not overfitted. One interesting contribution of BRNN is that the network allows evaluation of the uncertainty of a prediction.

***k*-Nearest Neighbors (*k*-NN).** Fundamental *k*-NN algorithms are fairly easy to understand as they are based on the hypothesis that similar compounds should present a similar activity [62]. Basically, a feature vector describes each compound of the training set. The whole set is mapped into a multidimensional feature space. The test set is then projected in the feature space and all distances between each compound of a test set and all the ones of the training set are computed. For the test set compounds, the predicted class is assigned by a majority vote in the *k* closest neighbors. The main challenge is then to choose the right way to calculate the distance and the right *k* value.

Even though *k*-NN is a popular method, we can only report two consistent studies on CYPs; one is included in a general overview of different methods (see Consensus models and multivariate analysis) and the other one is very recent. In 2007, Jensen *et al.* [63] used a Gaussian kernel weighted *k*-NN algorithm, a novel method based on Tanimoto similarity searches [64] on extended connectivity fingerprints (ECFP) and FCFP [65]. CYP2D6 and CYP3A4 inhibition was studied using 865 and 1037 compounds, respectively. *k* was fixed to 20 but, because of improvements in the method, other parameters had to be optimized such as a dynamic smoothing factor and an uncertainty term. In an unusual way, non-inhibitors were predicted with higher certainty. The external validation was successful with 83% for the training set, 82% for the external set regarding CYP2D6, and 87% and 88% regarding CYP3A4. Traditional *k*-NN was used on the CYP2D6 training set, again with *k* fixed to 20, and reached 81%. Thus, the Gaussian kernel weighted algorithm led to slightly better results than traditional *k*-NN. Finally, 14 other CYP2D6 external compounds were tested and 6 out of 14 were not classified, all belonging to medium inhibitors (*i.e.*, 40-60% inhibition). That weakness can be explained by the strong difference of chemical space covered by the training set and the test set.

Partial Least Squares (PLS) and Regressions. PLS is a very popular regression tool in the QSAR domain [66]; it was developed in the early 1980's by Wold [67]. In PLS, a linear model specifies the relationship between a dependent response, a biological activity for example, and a set of predictor variables such as molecular descriptors. In the early years, the reason of the success of the method was that it could extract useful correlations in cases where there were more variables than observations [68]. Strengths and weaknesses of the technique directly relate to the assumption that similar molecules have similar activities. PLS and QSAR in general are, thus, generally used to discover strong trends in a particular dataset but fail to depict slight but critical chemical variations. PLS studies on CYPs are numerous, that is why the results listed below cannot be considered as exhaustive.

In 1999, Ekins, one of the main contributor in the CYPs domain, and coworkers [69] proposed one of the first reliable PLS models with 3D and 4D QSAR study on CYP3A4 inhibitors. PLS was combined with molecular surface-weighted holistic invariant molecular descriptors (MS-WHIM) [70,71]. Those last ones consisted in a set of statistical parameters containing information about the size, shape, symmetry, and distribution of molecular surface point coordinates after weighted centering and PCA. It is very relevant as it suggests that the enzyme-ligand recognition arises along the molecular surface. The dataset contained only 11 CYP3A4 inhibitors. Results led to q^2 = 0.32 and q^2=0.44 when using multiple conformers. As a first PLS approach, the results were satisfying but could probably be improved in the following years.

A more consistent dataset of 53 CYP3A4 inhibitors and non-inhibitors was used by Wanchana *et al.* [72]. The descriptors used were 220 2D topological indices that were reduced to 20 by GA. The QSAR model obtained with the training set led to an *r* value of 0.88. A test set of 9 compounds was then quite successfully classified with an r_{pred} of 0.744 and a standard error of prediction (*s*) of 0.769. The authors pointed out that 2D topological descriptors were less computationally demanding than 3D descriptors and produced highly predictive models compared to 3D ones. That trend can be observed in many other studies.

More recently, Kriegl *et al.* [73] used an even larger dataset of 967 compounds with IC_{50} values measured on CYP3A4. The pool of descriptors was also large and diversified, *i.e.*, 32 2D descriptors, such as size, shape, lipophilicity, count of atoms and surface areas, 132 2D and 3D descriptors from MOE, 88 descriptors depicting interaction energies from the VolSurf package, and 68 descriptors based on AM1 quantum mechanical calculations. The authors performed a multivariate analysis based on PLS, PLS discriminant analysis (PLS-DA), and soft independent class modeling (SIMCA) to summarize the properties of strong and weak inhibitors. In PLS-DA, only the class-membership is taken into account. Therefore, during the regression, a matrix of two dummy variables, *i.e.*, 1 and 0 for a 2-classes problem, replaces the matrix of IC_{50}. Several models were built and PLS-DA seemed to slightly outperform the traditional PLS method. Indeed, PLS regression led to an r^2 value of 0.62 while PLS-DA led to an r^2 equal to 0.69. The external classification of 379 compounds was based on the definition of 3 classes: strong (IC_{50}<2μM), medium (2μM<IC_{50}<20μM), and weak (IC_{50}>20μM) inhibitors. The overall accuracy reached 66.0% for both PLS and PLS-DA. These results are acceptable for a 3-classes model; additionally, there was a low count of severe misclassifications (strong predicted as weak and *vice versa*). The strategy followed by the authors can be considered as really effective taking into account the large amount of data handled.

Other studies can also be briefly reported. Crivori and Poggesi [74] used PLS with grid-independent descriptors

(GRIND) to model 64 CYP2D6 inhibitors. The obtained r^2 value was 0.62 for the training set and a validation set was classified at 78%. Korhonen *et al.* [75] exploited comparative molecular field analysis (CoMFA) and obtained an r^2 value of 0.87 for CYP1A2 inhibitors. CoMFA was again used to predict CYP2D6 substrates, based on K_m values, by Haji-Momenian *et al.* [76] with an r^2 of 0.62 for 15 external compounds. Further PLS studies were reported for CYP2D6 inhibitors [77], CYP2B6 substrates [78], and CYP2C9 inhibitors [79].

Consensus Models and Multivariate Analysis. Consensus models are obtained by association of several models. The prediction is made by voting based on the single models or on their mean prediction. Some of the previously detailed studies presented some insights in the association of several methods to produce robust prediction models. We report two studies that explore extensively several methods and descriptors to predict CYPs inhibition

In 2005, Chohan *et al.* [80] studied the influence of four methods, PLS, MLR, CART, and Bayesian ANN (BNN), for the prediction of 109 CYP1A2 inhibitors. PLS used 17 of the 123 initial descriptors, and MLR, 5 of them. Classification and regression trees (CART) [81] were obtained with a consensus of 15 trees. BNN [82], which is less susceptible to overtraining/overfitting compared to classical ANN, produced a model with a 122-2-1 architecture on the basis of a *tanh* transfer function. For BNN only, the most relevant descriptors were selected by automatic relevance determination (ARD) [83]; leading to a reduction of the input nodes to 6. PLS, MLR, CART, and BNN all produced satisfying results with an r^2 value for the training set of 0.72, 0.71, 0.84, and 0.72, respectively, and a RMSE (root mean squared error, the error being one pIC_{50} unit) value of 1.0 for all models, with the exception of a value of 0.7 for CART. Consensus, *i.e.*, the average prediction from all four models, gave an RMSE of 0.84, which is a superior result compared to individual models, again with the exception of CART. For the test set of 68 molecules, the same conclusion could be drawn as the consensus gave an RMSE value of 1.2, that is better than PLS with RMSE = 1.3, MLR with RMSE = 1.4, and BNN with RMSE = 1.3, but not as good as CART with RMSE = 1.0. However, the models were not able to correctly predict the external test set, as the r^2 around the unity line was poor for all the models. Using a binary classification (pIC_{50} cutoff = 5), the global accuracy of the consensus model reached 83% but with an inhibitor classification rate of 56%; the test set was unbalanced toward non-inhibitors.

Arimoto *et al.* [84] continued the reasoning further for nearly 4000 CYP3A4 inhibitors by systematically exploring different descriptors and different methods such as RP, SVM, Bayesian classifiers, logistic regression, and *k*-NN. The descriptors were the 4096-bits Barnard chemical information (BCI) fingerprints [54], 166-bits MACCS keys [85], 13,608-bits typed graph triangle (TGT) fingerprint, and 156 topological and electrotopological indices calculated by MolconnZ [86]. For *k*-NN, *k* was equal to 9, and SVM used a radial basis kernel function. All combinations of methods and descriptors were explored and modeled. The very best model was obtained with SVM and MACCS keys with a coefficient of agreement κ value of 0.62 for 470 test compounds, followed by SVM and BCI with κ = 0.61, and RP and MolconnZ with κ = 0.60. The three best models were then used to build a voting consensus model. Requiring 1, 2 or 3 votes to be considered as inhibitors, consensus preformed with κ of 0.62 (Accuracy=81%), 0.65 (83%) and 0.57 (81%), respectively. The 2-votes method is then better than single classifiers.

The general conclusion about descriptors was that BCI fingerprints performed slightly better and TGT slightly worse. For the methods, Bayesian classifiers were significantly worse than the other methods and RP and SVM fairly better. It can be explained by the fact that Bayesian classifiers explicitly assume that descriptors are uncorrelated, which is rarely valid.

DISCUSSION

Strengths and Weaknesses of Algorithms. PLS combined to 3D QSAR descriptors is certainly the most controversial method because of its limits of applications when used for classification. It needs very accurate measures [52] and it is limited in size and structural diversity, *i.e.*, limited inside a family of compounds [34]. For the examples reviewed in this paper, the training datasets were composed of ~50 compounds. Moreover, compounds can bind in different modes, thus limits the application of 3D QSAR, especially for CYP3A4, a protein that has a relatively large binding pocket. Also, 2D QSAR reached comparable predictive abilities [72]. Nevertheless, it is one of the oldest methods, and, thus, one of the most used. In the CYPs domain, regression is less important than classification. Consequently QSAR models should be improved when used to predict a test set. Enhanced methods as PLS-DA used by Kriegl *et al.* [73] are certainly a valuable approach to smooth the weaknesses of QSAR models used for prediction.

ANN produced good prediction results for both CYPs inhibitors and substrates. Particularly, SOM are a powerful and fascinating tool to represent the topology of the chemical space covered by inhibitors, non-inhibitors substrates, non-substrates, and products of metabolism. The topology of the maps can be used as classifier with quite good results as in Korolev *et al.* [58] and Balakin *et al.* [59] studies. Another advantage of ANN is that some types of ANN allow the evaluation of the likely uncertainty of a prediction. Compared to standard least squares regressions, ANN are able to find a more general relationship between structure and activity.

For DTs and RP, the success rate is comparable to the best models obtained with the other methods, for example, ~90% accuracy for external validation [37,38]. Strength and weaknesses of RP can be explained by its simplicity. RP is easy to implement with a low number of parameters to define; RP is also rapid to screen large databases. But it needs particular attention to the quality of the datasets and descriptors. Indeed, RP may show poor predictability when derived from an excessively large pool of descriptors and can be very sensible to changes in the training set with cross-validation for example. That can be avoided by using ensemble strategies as suggested by Susnow *et al.* [29]. As a matter of fact, SVM is often superior to other methods in classification problems in many domains [12,68,87,88]. Moreover, it is not significantly affected by unbalanced datasets [89]. One can object that the method is sometimes computationally de-

Table 5. List and Definition of the Parameters Used by the Authors to Evaluate the Prediction Ability of their Models

Parameter	Definition	Range
Accuracy [84]	$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$	0 to 1 (often scaled in %)
C^2 (Generalized squared correlation coefficient) [20]	$C^2 = \frac{1}{n(k-1)} \sum_{i,j} \frac{(z_{ij} - e_{ij})^2}{e_{ij}}$ where k is the number of classes, z_{ij} are elements to be in class j while belonging to class i, and $e_{ij} = (1/n)\sum_j z_{ij} \sum_i z_{ij}$	0 to 1
Matthews correlation MCC [25]	Generalized correlation coefficient for k=2 (binary classification) or namely, $MCC = \frac{TP.TN - FN.FP}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}}$	-1 to 1
κ coefficient of agreement [57]	$\kappa = \frac{Accuracy - E}{1 - E}$ where $E = \frac{(TP + FN)(TP + FP) + (TN + FP)(TN + FN)}{(TP + TN + FP + FN)^2}$	0 to 1
RMSE	$RMSE = \sqrt{\frac{\sum_n (a_i - b_i)^2}{n}}$ where a_i are the correct values, b_i, the computed values and n, the total number of measures	∞ to 0
r_{pred}	$r_{pred} = \frac{n\sum_n a_i b_i - \sum_n a_i \sum_n b_i}{\sqrt{n\sum_n a_i^2 - (\sum_n a_i)^2}\sqrt{n\sum_n b_i^2 - (\sum_n b_i)^2}}$ where a_i are the correct values, b_i, the computed values and n, the total number of measures	-1 to 1
q^2	$q^2 = 1 - \frac{\sum_n (a_i - b_i)^2}{\sum_n (a_i - \bar{a}_i)^2}$ where a_i are the correct values, b_i, the computed values and n, the total number of measures	0 to 1
Sensitivity [84]	$Sensitivity = \frac{TP}{TP + FN}$	0 to 1 (often scaled in %)
Spearman's ρ rank [34]	$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$ where d_i is the difference between each rank of corresponding values of pairs, and n, the number of pairs of values.	-1 to 1

TP FP stands for the number of True Positives (number of compounds in class +1 c$FN$omputed in class +1), TN, True Negatives (number of compounds in class -1 computed in class -1),

manding and interpretation of the models is not easy and direct. A list of available software exploiting the different methods is provided in Table **6**.

Dataset Diversity. To increase the robustness of predictive models, highly diversified datasets are needed. However, the chemical space can be so vast that it is very difficult to collect and manage large training sets and obtain experimental measures. Frequently, readers are asked to trust the authors when they establish that their datasets are diversified. Nevertheless, different methods can be used to assess the diversity of datasets. A widespread method is to project the studied compounds in the chemical space of a large database to see the dispersal of the datasets; that can be done with PCA for example [38]. Yap and Chen [24] used the diversity index to quantify the similarity between all pairs of compounds [90]. Jensen *et al*. [63] performed a Jarvis-Patrick clustering on their training set to characterize the diversity. The more clusters, based, for example, on a Tanimoto coefficient of 0.85, the larger diversity. The authors obtained approximately 1.4 compounds per cluster, which corresponds to an acceptable diversity of the datasets. Actually, several methods can quantify the similarity between compounds within a dataset. Ideally that similarity has to be low to ensure the diversity but not too low to guarantee the extraction of general rules to build a model. Also one should remember that ML methods are often biased toward the larger class; that is why Eitrich *et al.* [22] developed dedicated methods to treat these cases (see the SVM section).

We can also point out the problem of data availability. Authors are generally requested to publish their datasets;

Table 6. Several Machine Learning Software Used in the Studies Reported in the Review

Machine Learning Method	Software	Website
SVM	LibSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvm/
SVM	SVM^{light}	http://svmlight.joachims.org/
Decision trees	Chemtree	http://www.goldenhelix.com/chemtreesoftware.html
Decision trees	MOE	http://www.chemcomp.com/
Decision trees	C4.5	http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html
Decision trees	S-PLUS	http://www.insightful.com/products/splus/default.asp
Neural networks and PLS	Cerius	http://www.accelrys.com/products/cerius2/
Neural networks	NeuroSolution	http://www.neurosolutions.com/
Decision trees and Bayesian theory	Pipeline Pilot	http://www.scitegic.com/
Neural networks	Stuttgart NN Simulator	http://www-ra.informatik.uni-tuebingen.de/SNNS/
k-Nearest neighbors and regressions	R statistical package	http://www.r-project.org/
PLS	GOLPE	http://www.miasrl.com/golpe.htm

sometime they base their studies on publicly available datasets [91,92]. It is very useful for the scientific community to have access to more and more data to improve the chemical space covered and, hence, the robustness of the predictions. It is also interesting to compare the models on the basis of benchmarking validation sets. In another hand, large private companies often publish really accurate studies based on unpublished in-house data, which is a bit contradictory with the previous point. Hopefully, the actual trend followed by most journals is to systematically publish the datasets used in the studies. Literature provides a good overview of the chemical diversity in the CYP field and can be profitably used to complement in house data. With such strategy the coverage of the chemical space can be improved. References [28,38,93] contain open source datasets for CYPs 3A4, 2D6, 1A2, and 2C9 substrates and inhibitors. Many more can be retrieved in recent articles.

Selection of Validation Sets. Another point concerning datasets is the choice of a relevant validation set. In most of the studies depicted herein, the authors do not explicitly define how they selected their external validation set. Datasets are split in training and test sets without any explanation. One of the most used methods is the *n*-fold cross validation which consist in dividing a dataset in *n* subsets. Each subset is used to validate a model built on the *n*-1 other subsets and it is repeated *n* times. If *n* equals the number of individuals in the dataset, the method is call leave-one-out cross validation. Unfortunately, chance is often the basis of the validation sets selection. The validation set can also come from a new collection of compounds which is presented to the predictive model without any relation with the training compounds. Generally, both training and validation sets are composed arbitrarily without particular care. Taking no caution is, however, quite dangerous and several authors often raise the argument of an irrelevant test set to explain the poor results of their validation. The point is that most of the ML methods applied in QSAR are suited for interpolation predictions but less for extrapolation. Therefore, it is crucial, to perform a reliable validation, to verify that the training and test sets cover the same range of descriptors, chemical space, and diversity. To treat this problem, a space filling algorithm, for example, was chosen by Kriegl *et al*. [20].

Class Definition. Classifying compounds in inhibitors/non-inhibitors or substrates/non-substrates can be a challenging problem. Ideally, the chosen threshold should have a biological meaning. Different hypothesis can be made, such as identifying substrates as non-inhibitors. Namely, threshold of K_i < 1 µM correspond to potent inhibitors causing DDI [94,95]. 10µM was used by Susnow *et al*. [29] because it corresponds to the expected blood concentration of typical drugs when administrated to therapeutic doses [96]; this value is confirmed by Chohan *et al*. [80], who defined that concentration as “problematic”. Another way to avoid that awkward question is to eliminate medium inhibitors and chose discriminating classes such as <1 µM and >50 µM. Nevertheless, the problem of class definition depends on the application of the model. Inhibitors, for example, should be clearly defined versus which CYP one is studying.

Descriptors Selection. Choosing the descriptors as input of an ML method is a delicate point. Once authors have selected the properties they want to depict, they generally try to minimize the number of descriptors used. Nevertheless, as underlined by Merkwirth *et al*. [26], aggressively restricted pools of features often lead to poor results. Hence, building a performing model should be a right balance between eliminating the noise (correlated descriptors, constant values, …) and being careful of not removing essential descriptors.

One can proceed manually as Yap and Chen [24], who tried different collections from 100 to 1000 features to observe that they reached a maximum of performance with a set of 300 descriptors. Really poor specificities were obtained with too much noise. In the presented studies, various methods such as GA [97], the McCabe algorithm [98], PCA [99], or Monte Carlo simulated annealing [100] were used to select only the relevant descriptors. Recursive feature elimination (RFE) was used by Xue *et al*. [101] to reduce the noise generated by too many descriptors. In RFE, when a SVM model is built, a weight is calculated for each feature. The features can then be ranked and the one with the smaller weight can be eliminated. The process can be repeated to eliminate more features. RFE is different from other methods of descriptor elimination as it is part of the model building and not an *a priori* statistic study based on the descriptors only. The application of RFE and SVM by Xue *et al*. was not on CYPs but still in the ADME thematic. In their work, RFE successfully reduced the initial set of 159 descriptors to sets of 22-31 descriptors, improving the prediction accuracy of their models. Finally, one would emphasize the particularly low number of descriptors used, as the consequence of the particular method used, *i.e.*, 9 for LWRP for example by Hudelson and Jones [37].

Key Descriptors. Any obtained model should be analyzed to figure out if it is physico-chemically relevant and to discover new information that can guide the drug design process. Key descriptors can be considered as those that have a high PCA or PLS coefficient, that appear often in the models, etc… In the case of CYPs binding, *i.e.*, inhibitors as well as substrates, three characteristics emerge from the existing models: lipophilicity, shape, and electrostatic interactions. Lipophilicity is mostly represented by *logP* and aromatic descriptors. It plays an important role in oxidation by CYPs [102]. High lipophilicity promotes CYP3A4 [73] and CYP1A2 inhibition [80], and it is certainly the case for other CYPs. As stated by Yap and Chen [24], CYP2D6 substrates contain generally a planar hydrophobic region, and for CYP2C9, aromaticity is often selected for model building. Molecular weight, size, topological indices, flexibility, and volume are enclosed in the shape descriptors and selected in almost all of the models studied. Large size of a molecule raises CYP3A4 inhibition; it can be explained by the higher stabilization through several hydrophobic interactions [103]. Electrostatic descriptors, such as polarizability, charge and surface charge, dipole moment, HOMO-LUMO gap, and electronegativities are also significant. Low polarity increases CYP3A4 inhibition; high charge, dipole moment, and HOMO-LUMO gap decreases CYP1A2 inhibition. Those features are also important to describe CYP1A2 substrates [104]. More interesting are the substructures or fragments that influence the behavior of a compound towards CYPs. Jensen *et al.* [63] stated that carboxyl acid is more frequent in non-inhibitors for both CYP2D6 and CYP1A2. Tertiary amine was frequent in CYP2D6 inhibitors and CYP3A4 non-inhibitors whereas phenol and sulfone were frequent in CYP2D6 non-inhibitors and CYP3A4 inhibitors. The number of hydroxyl groups is inversely proportional to CYP1A2 inhibition potency [105]. Let us note that it is accepted that nitrogen heterocycles bind strongly to the heme of CYPs, that CYP2D6 substrates contains usually a basic nitrogen [106], and that CYP3A4 substrates are neutral or basic [107].

CONCLUSIONS AND OUTLOOK

Predicting CYPs binding is a huge challenge due to the variety of compounds it can metabolize or that inhibit them. However, it is now a required task in the drug development to avoid unwanted drug-drug interactions and, hence, side effects. To screen large databases, rapid filters are needed and several machine learning methods can help to build them. The choice of the method depends mostly on the property of filter one wants to enhance: accuracy, speed, robustness, interpretability, … More and more, authors tune the original method to push farther the performances of filters, and they succeed as accuracy do not stop to raise. In the early days, good filters had 70% of accuracy, later, 80%, and nowadays, 90% is frequent. Improvements are not only dedicated to methods but also to the datasets, whose sizes are becoming larger, as well as the descriptors. Key descriptors bring additional information for the understanding of compounds interactions with CYPs. Interestingly, in the presented studies, 2D descriptors often outperform 3D ones, which is quite pleasing when speed of calculation is a limiting factor.

However, 3D studies of active site are absolutely needed in the next stages of CYPs binding comprehension. Structure based design, such as molecular docking [108-113], and 3D ligand modeling studies, such as pharmacophore elucidation [114-117], are other major topics in CYPs interaction. Three-dimensional modeling should seriously be taken into account as more and more CYPs crystallographic structures are available.

The future of *in silico* filters for CYPs is promising, as it seems to be more and more accurate with the emergence of dedicated metabolic descriptors [118], studies about compound stability regarding CYPs [119,120], and studies of regioselectivity of CYPs metabolism [121].

ACKNOWLEDGEMENTS

J.B. wishes to thank the Belgian “Fond pour la formation à la Recherche dans l’Industrie et dans l’Agriculture” (F.R.I.A.) for his Ph.D. fellowship. Authors would also like to acknowledge Prof. Ovidiu Ivanciuc and the referees for their pertinent remarks that helped to greatly improve the quality of the manuscript.

ABBREVIATIONS

ADME	=	Absorption, Distribution, Metabolism, Excretion
ANN	=	Artificial Neural Networks
ARD	=	Automatic Relevance Determination
BCI	=	Barnard Chemical Information
BNN	=	Bayesian Neural Network
BRNN	=	Bayesian Regularized Neural Network
CART	=	Classification And Regression Trees
CoMFA	=	Comparative Molecular Field Analysis
CYP	=	Cytochrome P450
DDI	=	Drug-Drug Interactions
DT	=	Decision Trees
ECFP	=	Extended Connectivity Fingerprint
FCFP	=	Functional Class Fingerprint
GA	=	Genetic Algorithm
GRIND	=	Grind Independent Descriptors
k-NN	=	*k*-Nearest Neighbors
LWRP	=	Line Walking Recursive Partitioning
MCC	=	Matthews Correlation Coefficient

ML	=	Machine Learning
MLR	=	Multiple Linear Regression
PCA	=	Principal Component Analysis
PLS	=	Partial Least Squares
PLS-DA	=	PLS Discriminant Analysis
PM-CSVM	=	Positive Majority Consensus Support Vector Machine
PP-CSVM	=	Positive Probability Consensus Support Vector Machine
QSAR	=	Quantitative Structure-Activity Relationship
RFE	=	Recursive Feature Elimination
RMSE	=	Root Mean Squared Error
RP	=	Recursive Partitioning
SEE	=	Standard Error of Estimation
SIMCA	=	Soft Independent Class Modeling
SOM	=	Self Organizing Maps
SEP	=	Standard Error of Prediction
SVM	=	Support Vector Machine
TGT	=	Typed Graph Triangle

REFERENCES

[1] Tredger, J. M.; Path, M. R. C.; Stoll, S. *Hosp. Pharmacist*, **2002**, *9*, 167-173.

[2] Yan, Z.; Caldwell, G. W. *Curr. Topics Med. Chem.*, **2001**, *1*, 403-425.

[3] de Groot, M. J. *Drug Discov. Today*, **2006**, *11*, 601-606.

[4] Wienkers, L. C.; Heath, T. G. *Nat. Rev. Drug Discov.*, **2005**, *4*, 825-833.

[5] Bidault, Y. *Expert Opin. Drug Metab. Toxicol.*, **2006**, *2*, 157-168.

[6] Guengerich, F. P. *AAPS J.*, **2006**, *8*, E101-E111.

[7] Wester, M. R.; Yano, J. K.; Schoch, G. A.; Yang, C.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.*, **2004**, *279*, 35630-35637.

[8] Yano, J. K.; Wester, M. R.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. *J. Biol. Chem.*, **2004**, *279*, 38091-38094.

[9] Rowland, P.; Blaney, F. E.; Smyth, M. G.; Jones, J. J.; Leydon, V. R.; Oxbrow, A. K.; Lewis, C. J.; Tennant, M. G.; Modi, S.; Eggleston, D. S.; Chenery, R. J.; Bridges, A. M. *J. Biol. Chem.*, **2006**, *281*, 7614-7622.

[10] Oprea, T. I.; Matter, H. *Curr Opin. Chem. Biol.*, **2004**, *8*, 349-358.

[11] Vapnik, V. In *Advances in Neural Information Processing Systems*; Moody, J. E., Hanson, S. J., Lippmann, R. P., Eds.; Morgan Kaufmann Publishers, Inc.: San Francisco, **1992**; *Vol. 4*.

[12] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.*, **2001**, *26*, 5-14.

[13] Jorissen, R. N.; Gilson, M. K. *J. Chem. Inf. Model.*, **2005**, *45*, 549-561.

[14] Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. *J. Chem. Inf. Model.*, **2005**, *45*, 1122-1133.

[15] Chen, B. N.; Harrison, R. F.; Pasupa, K.; Willett, P.; Wilton, D. J.; Wood, D. J.; Lewell, X. Q. *J. Chem. Inf. Model.*, **2006**, *46*, 478-486.

[16] Ivanciuc, O. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Cundari, T. R., Eds.; Wiley-VCH: Weinheim, **2007**, *23*, 291-400.

[17] Leong, M. K. *Chem. Res. Toxicol.*, **2007**, *20*, 217-226.

[18] Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge **2000**.

[19] Vapnik, V. *The Nature of Statistical Learning*; Springer: New York, **1995**.

[20] Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. *J. Comput.-Aided Mol. Des.*, **2005**, *19*, 189-201.

[21] www.chemcomp.com/.

[22] Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. *J. Chem. Inf. Model.*, **2007**, *47*, 92-103.

[23] Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. *J. Chem. Inf. Comput. Sci.*, **1994**, *34*, 109-116.

[24] Yap, C. W.; Chen, Y. Z. *J. Chem. Inf. Model.*, **2005**, *45*, 982-992.

[25] Matthews, B. W. *Biochim. Biophys. Acta*, **1975**, *405*, 442-451.

[26] Merkwirth, C.; Mauser, H. A.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1971-1978.

[27] Breiman, L.; Friedman, J.; Stone, C. J.; Olshen, R. A. *Classification and Regression Trees;* Chapman & Hall/CRC: London, **1984**.

[28] Zhang, H.; Singer, B. *Recursive Partitioning in the Health Sciences;* Springer: Berlin, **2005**.

[29] Susnow, R. G.; Dixon, S. L. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1308-1315.

[30] Rusinko, A.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 1017-1026.

[31] van Rhee, A. M.; Stocker, J.; Printzenhoff, D.; Creech, C.; Wagoner, P. K.; Spear, K. L. *J. Comb. Chem.*, **2001**, *3*, 267-277.

[32] Feng, J.; Lurati, L.; Ouyang, H.; Robinson, T.; Wang, Y. Y.; Yuan, S. L.; Young, S. S. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1463-1470.

[33] Breiman, L. *Mach. Learn.*, **2001**, *45*, 5-32.

[34] Ekins, S.; Berbaum, J.; Harrison, R. K. *Drug Metab. Dispos.*, **2003**, *31*, 1077-1080.

[35] www.goldenhelix.com/chemtreesoftware.html.

[36] Ekins, S. *Biochem. Soc. Trans.*, **2003**, *31*, 611-614.

[37] Hudelson, M. G.; Jones, J. P. *J. Med. Chem.*, **2006**, *49*, 4367-4373.

[38] Burton, J.; Ijjaali, I.; Barberan, O.; Petitet, F.; Vercauteren, D. P.; Michel, A. *J. Med. Chem.*, **2006**, *49*, 6231-6240.

[39] www.aureus-pharma.com.

[40] van Rhee, A. M. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 941-948.

[41] Buontempo, F. V.; Wang, X. Z.; Mwense, M.; Horan, N.; Young, A.; Osborn, D. *J. Chem. Inf. Model.*, **2005**, *45*, 904-912.

[42] Godden, J. W.; Furr, J. R.; Bajorath, J. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 182-188.

[43] DeLisle, R. K.; Dixon, S. L. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 862-870.

[44] Rosenblatt, F. *Psychol. Rev.*, **1958**, *65*, 386-408.

[45] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. *Nature*, **1986**, *323*, 533-536.

[46] Gallant, S. I. *IEEE Trans. Neural Netw.*, **1990**, *1*, 179-191.

[47] Kohonen, T. *Biol. Cybern.*, **1982**, *43*, 59-69.

[48] Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, **1999**.

[49] Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. *J. Chem. Inf. Model.*, **2006**, *46*, 1806-1818.

[50] Polley, M. J.; Burden, F. R.; Winkler, D. A. *Aust. J. Chem.*, **2005**, *58*, 859-863.

[51] Winkler, D. A. *Mol. Biotechnol.*, **2004**, *27*, 139-167.

[52] Molnar, L.; Keseru, G. M. *Bioorg. Med. Chem. Lett.*, **2002**, *12*, 419-421.

[53] O'Brien, S. E.; de Groot, M. J. *J. Med. Chem.*, **2005**, *48*, 1287-1291.

[54] Barnard, J. M.; Downs, G. M.; von Scholley-Pfab, A.; Brown, R. D. *J. Mol. Graph. Model.*, **2000**, *18*, 452-463.

[55] Kier, L. B.; Hall, L. H. *Pharm. Res.*, **1990**, *7*, 801-807.

[56] Rose, K.; Hall, L. H.; Kier, L. B. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 651-666.

[57] Cohen, J. *Educ. Psychol. Measure.*, **1960**, *20*, 37-46.

[58] Korolev, D.; Balakin, K. V.; Nikolsky, Y.; Kirillov, E.; Ivanenkov, Y. A.; Savchuk, N. P.; Ivashchenko, A. A.; Nikolskaya, T. *J. Med. Chem.*, **2003**, *46*, 3631-3643.

[59] Balakin, K. V.; Ekins, S.; Bugrim, A.; Ivanenkov, Y. A.; Korolev, D.; Nikolsky, Y. V.; Skorenko, A. V.; Ivashchenko, A. A.; Savchuk, N. P.; Nikolskaya, T. *Drug Metab. Dispos.*, **2004**, *32*, 1183-1189.

[60] Wang, Y. H.; Li, Y.; Li, Y. H.; Yang, S. L.; Yang, L. *Bioorg. Med. Chem. Lett.*, **2005**, *15*, 4076-4084.

[61] Mackay, D. J. C. *Network Comp. Neural Syst.*, **1995**, *6*, 469-505.

[62] Shakhnarovich, G.; Indyk, P.; Darrell, T. *Nearest-Neighbor Methods in Learning And Vision: Theory And Practice*; MIT Press: Cambridge (USA), **2006**.

[63] Jensen, B. F.; Vind, C.; Padkjaer, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. *J. Med. Chem.*, **2007**, *50*, 501-511.

[64] Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.

[65] Rogers, D.; Brown, R. D.; Hahn, M. *J. Biomol. Screen.*, **2005**, *10*, 682-686.

[66] Wold, S.; Johansson, E.; Cocchi, M. In *3D-QSAR in Drug Design: Theory, Methods, and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, **1993**, pp. 523-550.

[67] Wold, H. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; Wiley: New York, **1985**, Vol. *6*, pp. 581-591.

[68] Oprea, T. I. *Chemoinformatics in Drug Discovery*; Wiley-VCH: Weinheim, **2005**.

[69] Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. *J. Pharmacol. Exp. Ther.*, **1999**, *290*, 429-438.

[70] Bravi, G.; Wikel, J. H. *Quant. Struct.-Act. Rel.*, **2000**, *19*, 29-38.

[71] Bravi, G.; Wikel, J. H. *Quant. Struct.-Act. Rel.*, **2000**, *19*, 39-49.

[72] Wanchana, S.; Yamashita, F.; Hashida, M. *Pharm. Res.*, **2003**, *20*, 1401-1408.

[73] Kriegl, J. M.; Eriksson, L.; Arnhold, T.; Beck, B.; Johansson, E.; Fox, T. *Eur. J. Pharm. Sci.*, **2005**, *24*, 451-463.

[74] Crivori, P.; Poggesi, I. *Basic Clin. Pharmacol.*, **2005**, *96*, 251-253.

[75] Korhonen, L. E.; Rahnasto, M.; Mahonen, N. J.; Wittekindt, C.; Poso, A.; Juvonen, R. O.; Raunio, H. *J. Med. Chem.*, **2005**, *48*, 3808-3815.

[76] Haji-Momenian, S.; Rieger, J. M.; Macdonald, T. L.; Brown, M. L. *Bioorg. Med. Chem.*, **2003**, *11*, 5545-5554.

[77] Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. *Pharmacogenetics*, **1999**, *9*, 477-489.

[78] Ekins, S.; Bravi, G.; Ring, B. J.; Gillespie, T. A.; Gillespie, J. S.; Vandenbranden, M.; Wrighton, S. A.; Wikel, J. H. *J. Pharmacol. Exp. Ther.*, **1999**, *288*, 21-29.

[79] Jones, J. P.; He, M. X.; Trager, W. F.; Rettie, A. E. *Drug Metab. Dispos.*, **1996**, *24*, 1-6.

[80] Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. *J. Med. Chem.*, **2005**, *48*, 5154-5161.

[81] Steinberg, D.; Colla, P. *Tree-Structured Non-Parametric Data Analysis*; Salford Systems: San Diego, **1995**.

[82] Neal, R. M. *Bayesian Learning for Neural Networks*; Springer-Verlag: New York, **1996**.

[83] Burden, F. R.; Ford, M. G.; Whitley, D. C.; Winkler, D. A. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1423-1430.

[84] Arimoto, R.; Prasad, M. A.; Gifford, E. M. *J. Biomol. Screen.*, **2005**, *10*, 197-205.

[85] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 1273-1280.

[86] www.edusoft-lc.com/molconn/.

[87] Czerminski, R.; Yasri, A.; Hartsough, D. *Quant. Struct.-Act. Rel.*, **2001**, *20*, 227-240.

[88] Meyer, D.; Leisch, F.; Hornik, K. *Neurocomputing*, **2003**, *55*, 169-186.

[89] Cai, C. Z.; Han, L. Y.; Ji, Z. L.; Chen, X.; Chen, Y. Z. *Nucl. Acids Res.*, **2003**, *31*, 3692-3697.

[90] Perez, J. J. *Chem. Soc. Rev.*, **2005**, *34*, 143-152.

[91] Rendic, S.; DiCarlo, F. J. *Drug Metab. Rev.*, **1997**, *29*, 413-580.

[92] Rendic, S. *Drug Metab. Rev.*, **2002**, *34*, 83-448.

[93] Terfloth, L.; Bienfait, B.; Gasteiger, J. *J. Chem. Inf. Model.*, **2007**, *47*, 1688-1701.

[94] Lin, J. H.; Pearson, P. G. In *Drug-Drug Interactions*; A.D., R., Ed.; Marcel Dekker: New York, **2002**, p 415-438.

[95] Blanchard, N.; Richert, L.; Coassolo, P.; Lave, T. *Curr. Drug Metab.*, **2004**, *5*, 147-156.

[96] Hamelin, B. A.; Bouayad, A.; Drolet, B.; Gravel, A.; Turgeon, J. *Drug Metab. Dispos.*, **1998**, *26*, 536-539.

[97] Mitchell, M. *An Introduction to Genetic Algorithms* The MIT Press: Cambridge (USA), **1998**.

[98] McCabe, G. P. *Technometrics*, **1984**, *26*, 137-144.

[99] Jolliffe, I. T. *Principal Component Analysis;* Springer: Berlin, **2002**.

[100] Liu, J. S. *Monte Carlo Strategies in Scientific Computing;* Springer: Berlin, **2002**.

[101] Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. *J. Chem. Inf. Comput. Sci.*, **2004**, *44*, 1630-1638.

[102] Hansch, C. *Drug Metab. Rev.*, **1972**, *1*, 1-14.

[103] Szklarz, G. D.; Halpert, J. R. *J. Comput.-Aided Mol. Des.*, **1997**, *11*, 265-272.

[104] Lewis, D. F. V. *Biochem. Pharmacol.*, **2000**, *60*, 293-306.

[105] Lee, H.; Yeom, H.; Kim, Y. G.; Yoon, C. N.; Jin, C. B.; Choi, J. S.; Kim, B. R.; Kim, D. H. *Biochem. Pharmacol.*, **1998**, *55*, 1369-1375.

[106] Langowski, J.; Long, A. *Adv. Drug Deliv. Rev.*, **2002**, *54*, 407-415.

[107] Smith, D. A.; Ackland, M. J.; Jones, B. C. *Drug Discov. Today*, **1997**, *2*, 479-486.

[108] Keseru, G. M. *J. Comput. Aid. Mol. Des.*, **2001**, *15*, 649-657.

[109] Kemp, C. A.; Flanagan, J. U.; van Eldik, A. J.; Marechal, J. D.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. *J. Med. Chem.*, **2004**, *47*, 5340-5346.

[110] Fukunishi, Y.; Hojo, S.; Nakamura, H. *J. Chem. Inf. Model.*, **2006**, *46*, 2610-2622.

[111] de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. *J. Med. Chem.*, **2006**, *49*, 2417-2430.

[112] Bazeley, P. S.; Prithivi, S.; Struble, C. A.; Povinelli, R. J.; Sem, D. S. *J. Chem. Inf. Model.*, **2006**, *46*, 2698-2708.

[113] Marechal, J. D.; Yu, J. L.; Brown, S.; Kapelioukh, I.; Rankin, E. M.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. *Drug Metab. Dispos.*, **2006**, *34*, 534-538.

[114] de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. *J. Med. Chem.*, **1999**, *42*, 4062-4070.

[115] Ekins, S.; Bravi, G.; Wikel, J. H.; Wrighton, S. A. *J. Pharmacol. Exp. Ther.*, **1999**, *291*, 424-433.

[116] Schuster, D.; Laggner, C.; Steindl, T. M.; Palusczak, A.; Hartmann, R. W.; Langer, T. *J. Chem. Inf. Model.*, **2006**, *46*, 1301-1311.

[117] Mao, B.; Gozalbes, R.; Barbosa, F.; Migeon, J.; Merrick, S.; Kamm, K.; Wong, E.; Costales, C.; Shi, W.; Wu, C.; Froloff, N. *J. Chem. Inf. Model.*, **2006**, *46*, 2125-2134.

[118] Keseru, G. M.; Molnar, L. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 437-444.

[119] Crivori, P.; Zamora, I.; Speed, B.; Orrenius, C.; Poggesi, I. *J. Comput.-Aided Mol. Des.*, **2004**, *18*, 155-166.

[120] Sciabola, S.; Morao, I.; de Groot, M. J. *J. Chem. Inf. Model.*, **2007**, *47*, 76-84.

[121] Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. *J. Med. Chem.*, **2005**, *48*, 6970-6979.

[122] http://medicine.iupui.edu/flockhart/table.htm.

Received: May 8, 2008 **Revised: July 2, 2008** **Accepted: August 29, 2008**