JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

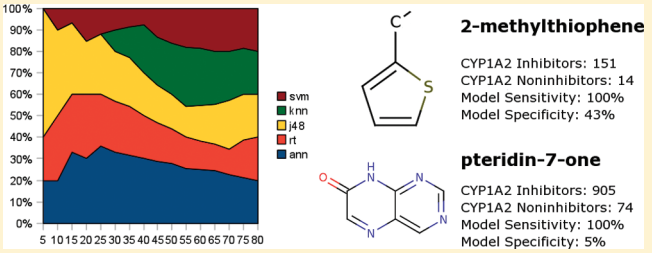
A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition

Sergii Novotarskyi,[†,‡] Iurii Sushko,[†,‡] Robert Körner,[†,‡] Anil Kumar Pandey,[‡] and Igor V. Tetko[†,‡,*]

[†]eADMET GmbH, Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

[‡]Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum Muenchen, German Research Center for Environmental Health (GmbH), Ingolstädter Landstrasse 1, D-85764 Neuherberg, Germany

ABSTRACT: Prediction of CYP450 inhibition activity of small molecules poses an important task due to high risk of drug−drug interactions. CYP1A2 is an important member of CYP450 superfamily and accounts for 15% of total CYP450 presence in human liver.



This article compares 80 in-silico QSAR models that were created by following the same procedure with different combinations of descriptors and machine learning methods. The training and test sets consist of 3745 and 3741 inhibitors and noninhibitors from PubChem BioAssay database. A heterogeneous external test set of 160 inhibitors was collected from literature. The studied descriptor sets involve E-state, Dragon and ISIDA SMF descriptors. Machine learning methods involve Associative Neural Networks (ASNN), K Nearest Neighbors (kNN), Random Tree (RT), C4.5 Tree (J48), and Support Vector Machines (SVM). The influence of descriptor selection on model accuracy was studied. The benefits of "bagging" modeling approach were shown.

Applicability domain approach was successfully applied in this study and ways of increasing model accuracy through use of applicability domain measures were demonstrated as well as fragment-based model interpretation was performed.

The most accurate models in this study achieved values of 83% and 68% correctly classified instances on the internal and external test sets, respectively. The applicability domain approach allowed increasing the prediction accuracy to 90% for 78% of the internal and 17% of the external test sets, respectively. The most accurate models are available online at http://ochem.eu/models/Q5747.

INTRODUCTION

The prediction of metabolism of molecules is of great interest for drug discovery. Cytochromes P450 (CYP450) are a superfamily of enzymes, involved in metabolism of a large number of xenobiotic compounds.[1] CYP450 are involved in metabolism of a large amount of drugs, currently present on the market.[2] Individual CYP enzymes in families 1, 2, and 3 metabolize xenobiotics, including the majority of small molecule drugs currently in use.[3] The distinctive feature of CYP450 enzymes is broad and overlapping substrate specificity.[4] Approximately 70% of currently used drugs are cleared through metabolism and ten CYP450 forms in human liver carry out virtually the whole CYP-mediated metabolism. It is worth noting that most drugs, which are cleared by the CYP system, are metabolized through several CYP forms. As a general rule, drugs that are metabolized by a single CYP form are more susceptible to drug interactions than drugs metabolized by multiple forms.

The promiscuity with respect to substrates makes the CYP450 prone to inhibition by a large amount of drugs, which may lead to clinically significant drug−drug interactions.[5,6] Similarly to a large number of other proteins, CYP450 enzymes are prone to both competitive and noncompetitive inhibition. In competitive inhibition, there is competition between the substrate and inhibitor to bind to the same position on the active site of the enzyme. In the noncompetitive mode of inhibition, the active binding site of the substrate and inhibitor is different from each other. In the case of noncompetitive inhibition, the inhibitor binds to the enzyme−substrate complex, but not to the free enzyme entity. In practice, mixed-type inhibition displaying elements of both competitive and noncompetitive inhibition are frequently observed for CYP450 enzymes.

CYP450 inhibition can lead to decreased elimination of compounds dependent on metabolism for systemic clearance. If a drug is metabolized mainly via a single pathway, CYP inhibition may result in an increased steady-state concentration and accumulation ratio and nonlinear kinetics as a consequence of the saturation of enzymatic processes. Especially with prodrugs, inhibition may result in a decrease in the amount of the active drug form. Thus, inhibition of CYP may lead to toxicity or lack of efficacy of drugs.[3] Therefore, early prediction of CYP450-related activity of compounds may help to avoid the pursuit of drug candidates with these undesirable effects.

CYP1A2 is a major enzyme in the metabolism of a number of important chemicals, which typically belong structurally to the group of planar polyaromatic amides and amines.[7] It accounts for 15% of total CYP contents in human liver and is responsible for the metabolizm of approximately 5% of therapeutically used drugs.[8,9] Amitriptyline, ethoxyresorufin, caffeine, fluvoxamine, phenacetin, theophylline, clozapine, melatonin, haloperidol,

Received: February 22, 2011
Published: May 20, 2011

zolmitriptan and tizanidine are biotransformed predominantly by CYP1A2.[10] CYP1A2, its partisipation in xenobiotics metabolism and corresponding implications for drug development is an intensivly studied topic in medicinal chemistry.[11]

Computational methods (including QSAR methods) are especially attractive in the early stages of drug discovery since they can be used for screening of virtual molecular libraries, resulting in a dramatic decrease of potential candidates. The successful use of QSAR methods for prediction of CYP450 (and in particular CYP1A2) activity was shown.[12−17] However, the previous studies were limited with respect to the number of applied machine learning methods and diversity of descriptors as well as the lack of a common approach to model evaluation and the use of applicability domain methods.

In this article, we compared the performance of several systematically developed QSAR models for CYP1A2 inhibition. We built the models with a range of machine learning methods on a variety of descriptor sets. Our goal was to understand how the accuracy of prediction of CYP1A2 inhibitors depends on the different machine learning methods and descriptor sets, in particular descriptor selection, and to find the combination of descriptors and machine learning methods that would yield the highest predictivity. We also studied the influence of ensemble and bagging approaches, as well as descriptor selection approaches on resulting model predictivity. A fragment-based approach to model interpretation was used to reveal several fragments, possibly relevant to inhibitory activity. The benefits of characterization of models′ prediction accuracies were demonstrated for an application of the developed models to the external heterogeneous data set collected from literature.

Another important goal of this study was to provide publicly accessible models that could be easily used by chemoinfomatics community to screen their compounds for CYP1A2 inhibition. While there were many publications in this area, in most cases the published models and data are not publicly available and can not be used/tested by the community. The models developed in this article are publicly accessible at http://ochem.eu/models/Q5747 and can be used online. Moreover, the use of these models will allow to better evaluate the usefulness of HTS screening techniques and *in silico* approaches for idetification of CYP inhibitors.

DATA

PubChem data set. The structures and the inhibitor/noninhibitor labels for the compounds were collected from PubChem BioAssay database for human CYP1A2 inhibition assay with internal PubMed ID of AID410.[18] The description of the BioAssay experiment shows that the demethylation of luciferin 6′ methyl ether (Luciferin-ME; Promega-Glo) to luciferin was used as a target reaction for human CYP1A2 for this data set. The luciferin was then measured by luminescence after the addition of a luciferase detection reagent. The data set obtained from PubChem contained 8348 compounds, out of which 4175 were determined as active, 3673 − inactive, 713 − inconclusive. The protocol summary of the assay is available from the assay page on PubChem. The detailed protocol description is available in the Promega-Glo technical bulletin.[19]

Prior to any further analysis, all molecules were processed by Chemaxon standardizer[20] and, if required, were dearomatized to the Kekule representation. All the nitrogroups were converted to a consistent representation. The molecules were neutralized and all counterions and salts were removed. This procedure produced a number of duplicates, which were determined using InChI keys. If the same molecule was in both "active" and "inactive" sets or if a molecule was found in an "inconclusive" set, as specified by PubChem, it was removed from all sets. This was the case for 241 molecules. The number of nonconflicting inconclusives was 543 compounds. There were also 66 molecules, that were duplicates within "inactive" or "active" lists, respectively. As a result of this preprocessing a nonredundant set of 4016 active and 3470 inactive molecules (a total of 7486 molecules) was formed.

Experimental accuracy of the PubChem data set. To have a basis for comparison of model accuracy to the experimental accuracy, we considered the inconclusive compounds in the data set as experimental errors. Under these assumptions the experimental error of the data set is 713/8348 = 0.085−9%. Since not all inconclusive compounds should be treated as experimental errors, this value is an overestimation. However, it provides a lower boundary for accuracy estimation.

Test set of published CYP1A2 inhibitors. In addition to the PubChem assay, a test data set with molecules collected from literature was used to validate the accuracy of the models. The compounds were introduced from a review article of human CYP metabolism data.[1] This article reported 160 CYP1A2 inhibitors, collected from over 100 different sources (for some molecules the values were confirmed in several articles). The inhibitors were measured using several protocols and different sample drugs and thus were more diverse as compared to the data that was used to develop models in the current study. The molecules were preprocessed similarly to the PubChem data.

METHODS

Descriptors calculation. One of the goals of the study was to determine the influence of different representation of molecules on the quality of models for the CYP1A2 inhibitor prediction. Three descriptors sets were used: fragments-base descriptors (ISIDA SMF),[21,22] 2D topological descriptors (E-state)[23,24] and a diverse set of 0D − 3D descriptors (Dragon).[25] Below we describe these descriptors in more detail.

ISIDA SMF descriptors were calculated using the fragmentation tool from the ISIDA suite.[21] The *substructural molecular fragments* (SMF) method is based on the splitting of a molecule into fragments. The fragment type is then a descriptor, and the number of occurrences of this fragment in a molecule is the value for this descriptor. Two different types of fragments are considered: "sequences" and "augmented atoms". For each type of fragment three subtypes can be defined **AB** (atom and bond types), **A** (atom types only), and **B** (bond types only). In this study the **AB** type descriptors were used. These descriptors were shown to provide better performance compared to other two sets in several previous studies.[21] For the sequences, the length of calculated fragments was limited to between 2 and 5 atoms. This resulted into calculation of 3534 different descriptors for the described data set.

Atom type E-state indices and molecular bond E-state indices were calculated using a custom written tool that implements the procedures described in appropriate articles by Hall and Kier.[24] These descriptors combine electronic and topological properties of the described molecules. Each atom in the molecular graph is represented by an E-state variable, which encodes the intrinsic electronic state of the atom as perturbed by the

electronic influence of all other atoms in the molecule within the context of the topological character of the molecule. The E-state index for an atom or bond consists of an intrinsic value for that atom/bond plus a term for its perturbation by all the other atoms in the molecule. For every atom type and bond type in the molecule the calculated indices are summed. The total amount of E-state indices for this data set was 425.

Dragon[25] is a software tool licensed by Talete inc. The Linux version of Dragon — dragonX 1.2.4, which calculates 1664 molecular descriptors, was used. These descriptors cover 0D - 3D descriptors which are arranged into 20 blocks. The 0D descriptors are the descriptors independent of any knowledge concerning the molecular structure. Examples of 0D descriptors are total atom number, absolute or relative number of specific atom types, absolute or relative number of specific bond types, etc. The 1D descriptors are calculated over such one-dimensional representations of a molecule as lists of fragments or functional groups of interest present in the molecule. The 2D descriptors are derived from two-dimensional topological representation of the molecule and include topological information indices, molecular profiles and 2D autocorrelation descriptors. The 3D descriptors are base on a three-dimensional representation of the molecule and include WHIM, GETAWAY and 3D-MoRSE descriptors.[25]

The total amount of descriptors produced by all three tools was 5623.

3D structure generation. Since some of the used descriptors require valid 3D structures, the Corina[26] software was used to generate 3D conformations of molecules from their 2D representations obtained on the previous step. All compounds were converted to 3D structures without errors. The version of Corina used in this study is 3.44 (14.05.2008). Corina tool by was chosen for 3D conformation generation due to its high conversion rates, ability to handle wide variety of atom types, and the reported ability to generate conformations close to those obtained by X-ray measurements.[27] Corina was successfully used before in other CYP450-related studies.[28]

Selection of training and test sets. The initial PubChem data set was randomly split into two subsets containing 3745 and 3741 records each. The first set was used as a training set, the other one as a test set. Only the training set was used for model development for all studies reported in this article. The test set was used to get an unbiased estimation of prediction ability of developed models. The training set contained 2014 inhibitors and 1731 noninhibitor. The test set contained 2003 inhibitors and 1738 noninhibitors.

Selection of descriptors. The models in this study were created both with full set of descriptors and with the use of descriptor selection procedure. The selection of the nonredundant set of descriptors was performed using only the training set. The descriptors were filtered by a "best first" search method using "correlation-based feature subset selection" method as an attribute evaluator, which is implemented in the Weka[29] software. The "best first" method searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility. The "correlation-based feature subset selection" method evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred.[30] The mentioned method of descriptor selection was found to be efficient in other QSPR studies.[16]

Machine learning methods. Several popular machine-learning methods that were found efficient for QSAR modeling were used in this study. When applied to the same data sets, these methods provide a basis for comparison of efficiency of each method to predict CYP1A2 inhibitors. The analyzed methods were Associative Neural Networks (ASNN),[31,32] k Nearest Neighbors (kNN), Random Tree (RT),[33] C4.5 Tree (J48),[34] and Support Vector Machines (SVM)[35] as implemented in LibSVM.[36]

kNN method predicts the target activity as an average value of activities of its k nearest neighbors. The comparison is performed in space of descriptors and Euclidian distance was used.

ASNN is a combination of an ensemble of feed-forward neural networks and the kNN. This method's main feature is neural network ensemble bias correction achieved by the kNN method. The metrics used for kNN is correlation between the responses of individual neural networks in an ensemble. Therefore, the corrections are performed in space of ensemble residuals.

RT is a Weka[29] implementation of the random decision tree algorithm. It is a decision tree with no pruning and considering only $\log_2(N)$ of descriptors in each node (where N is a total amount of available descriptors).

J48 is a Weka implementation of the C4.5 pruned decision tree. It tries to recursively partition the data set into subsets by evaluating the normalized information gain (difference in entropy) resulting from choosing a descriptor for splitting the data. The descriptor with the highest information gain is used on every step. The training process stops when the resulting nodes contain instances of single classes or if no descriptor can be found that would result to the information gain.

SVM used in this study is a kernel-based classification method. In this method the input variables are mapped to a higher dimensional space by the use of a kernel function and are then classified by constructing a hyperplane in this space. In this study, the radial basis function kernel was used.

All the methods were also used in a combination with the bagging approach.[37] In each bagging session, 100 models were developed. For each model a training set was obtained from the original training set by resampling with replacement (each generated replica was of the size of the original set, and was created by randomly choosing entries, duplicates were allowed). The result of each classification was determined by voting among 100 models over the classified sample. Mean values and standard deviations of model predictions over each class for a specific instance were used to evaluate the confidence of predictions. This approach is discussed in the applicability domain section of this paper.

Prediction quality assessment. For binary classification models, there is a number of established metrics that identify model performance. In the definitions below numbers of true positives, true negatives, false positives, and false negatives are denoted as *TP, TN, FP, FN*, respectively.

In this study, the accuracy

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}$$

was used as the main measure of model performance in all statistical tests.

The weighted accuracy

$$WACC = 0.5\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

and Matthew's correlation coefficient

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

were also calculated to characterize the predictivity of the studied models.

The bootstrap statistical test[38] was used to estimate the significance of the differences in model performance. For this test, 10000 of replicas were generated by resampling with replacement from the analyzed set. The 0.05 confidence value was chosen, i.e. conclusion of significant differences in models performance was made if one model performed better than the other on more than 95% of the replicas. All the models were then compared in a pairwise manner.

Applicability domain assessment. The developed models do not have the same performance for all possible chemicals. Thus, it is very important to distinguish reliable and nonreliable predictions: the former predictions can be used in place of experimental measurements while latter ones should be tested in experiments. This can be done using "applicability domain" techniques. AD methods rely on finding a measure that correlates with the accuracy of predictions (this measure is referred to as distance to model, DM[39]). This way it is possible to predict the model accuracy for a particular compound and select a subset of most confident predictions.

In this study we applied the recently published "distance to model" - PROB-STD[39,40] DM, which can be easily, calculated using bagging and ensemble approaches. This measure was successfully used in our preliminary study of CYP450[39] set as well as it is analyzed in details using AMES challenge set.[40]

It is defined as

$$d_{PROB-STD}(J) = \min\left\{ \begin{array}{l} \int_0^{+\infty} N(x, y(J), d_{STD}(J))dx \\ \int_{-\infty}^{0} N(x, y(J), d_{STD}(J))dx \end{array} \right\}$$

where $y(J)$ is a quantitative value of prediction for compound J, $d_{STD}(J)$ is the standard deviation of predictions for this compound, $N(x,y(J),d_{STD}(J))$ is the normal distribution density function with mean $y(J)$ and standard deviation $d_{STD}(J)$. The $y(J)$ and $d_{STD}(J)$ are calculated over a set of predictions for a molecule in an ensemble or bag of individual models.

This distance to model has a natural interpretation. For a compound classified as $\{+1\}$, the square of $d_{PROB-STD}(J)$ will correspond to the probability of compound's classification to the opposite class, i.e. $\{-1\}$ and vise versa.

Model interpretation. The significant number of descriptors and the nonlinear character of all the machine learning methods used in this study make the descriptor-based approach to model interpretation infeasible. Additinally, descriptor-based approach is less useful for decision-support purposes in drug candidate optimization, as structural changes to the molecule cause the change of the whole set of related descriptors. A fragment-based approach was adopted instead, where model predictions are grouped in a fragment-based manner.

Among over 2 million possible molecular subfragments of our data set we selected around 15 thousand that appeared in at least 10 molecules in a data set. For each fragment, we calculated the number of inhibitors and noninhibitors containing this fragment, as well as the number of correctly and incorrectly classified instances containing this fragment.

Based on that a specific group of fragments of interest was separated: fragments specific for inhibitors (ratio of inhibitor molecules containing these fragments to noninhibitor molecules more than ten). This group was analyzed in terms of model performance.

Model availability. The study was performed with the use of Online Chemical Modeling Environment (OCHEM) - http://ochem.eu. The OCHEM is a web-based platform that aims to automate and simplify the typical steps required for QSAR modeling and provides facilities for collaboration in QSAR studies, exchange of data and for publication of results.[41] It allows to store and share QSAR data sets and models and to compare results and approaches.

Models published at OCHEM are easily reproducible - the training and test sets are public, the model building workflow is described in details, and all the tools involved in model creation (molecule preprocessing tools, descriptor calculation and filtering programs, machine learning methods) are available. Model description also includes all the parameters for every node of the model building workflow.

Some of the best performing models in this study along with the data used for their creation are freely and publicly available for verification and usage at http://ochem.eu/models/Q5747.

RESULTS AND DISCUSSION

The main goal of this work was to compare the accuracy of the CYP1A2 classification models built with different machine learning methods on different sets of descriptors.

There were 4 parameters that were benchmarked in this study:

- machine learning method (ASNN, KNN, RT, J48, SVM)
- ensemble approach (single model, ensemble (for ASNN)/ bagging (for other methods))
- descriptor set (E-state, SMF, Dragon, Full set)
- descriptor selection protocol (Full set, Selected set)

Given all possible combinations, the total of 80 different models were generated and compared. Table 1 displays 20 top performing models out of these 80. The selection was performed based on overall model accuracy (ACC). The table contains the details of each model (descriptor set, machine learning method, ensemble approach) as well as additional accuracy measures - weighted accuracy (WACC) and Matthew's correlation coefficient (MCC). The table is divided into three groups. Within each group the models are statistically nonsignificantly different to the top model in the group with the significance level of 0.05.

To investigate the influence of the studied parameters of models on their test set accuracy, the cumulative plots were built. First, the models were sorted according to ACC in descending order and n top-performing models were selected. Second, among the list of n-top ranked models the percentage of models of each particular machine learning method was calculated. Figure 1a shows the calculated percentage of models, built by a particular machine learning method (y axis) among the n top-performing models (x axis). Number n changes from 5 to 80 with a step of 5 models. Methods with higher areas in the left part of the plot had higher performance.

Figure 1b uses the same concept to illustrate the difference in descriptor performance, while figure 1c shows the performance of ensemble versus single method.

Table 1. The performance of best 20 models for the internal test set of 3741 of CYP1A2 inhibitors and noninhibitors from PubChem BioAssay database

	DESCR	SEL	METHOD	ENSEMBLE	ACC	WACC	MCC	SENS	SPEC
1	All	no	ASNN	ensemble	0.827	*0.827*	*0.653*	*0.827*	*0.827*
2	All	no	J48	bagging	0.827	*0.827*	*0.653*	*0.827*	*0.827*
3	All	yes	J48	bagging	0.823	*0.823*	*0.645*	*0.823*	*0.823*
4	Dragon	no	J48	bagging	0.820	*0.821*	*0.640*	*0.807*	*0.835*
5	Dragon	yes	RT	bagging	0.820	*0.819*	*0.638*	*0.833*	*0.804*
6	All	yes	RT	bagging	0.820	*0.818*	*0.637*	*0.846*	*0.789*
7	SMF	no	J48	bagging	0.819	*0.819*	*0.637*	*0.819*	*0.819*
8	Dragon	no	ASNN	ensemble	0.818	*0.819*	*0.636*	*0.804*	*0.833*
9	SMF	no	SVM	single	0.818	*0.818*	*0.635*	*0.818*	*0.818*
10	E-state	no	RT	bagging	0.817	*0.818*	*0.634*	*0.803*	*0.832*
11	Dragon	yes	J48	bagging	0.814	*0.814*	*0.627*	*0.814*	*0.814*
12	All	yes	ASNN	ensemble	0.813	*0.813*	*0.625*	*0.813*	*0.813*
13	SMF	no	ASNN	ensemble	0.812	*0.813*	*0.624*	*0.798*	*0.827*
14	Dragon	no	ASNN	single	0.811	*0.816*	*0.633*	*0.745*	*0.886*
15	All	no	RT	bagging	0.811	*0.811*	*0.621*	*0.811*	*0.811*
16	E-state	no	ASNN	ensemble	0.810	*0.810*	*0.619*	*0.810*	*0.810*
17	E-state	no	SVM	bagging	0.810	*0.812*	*0.622*	*0.783*	*0.840*
18	E-state	no	SVM	single	0.809	*0.812*	*0.622*	*0.769*	*0.854*
19	Dragon	no	RT	bagging	0.809	*0.808*	*0.616*	*0.822*	*0.793*
20	SMF	no	RT	bagging	0.808	*0.807*	*0.614*	*0.821*	*0.792*

ASNN — Associative Neural Networks,[31,32] RT and J48 — random trees and C4.5 pruned trees as implemented in WEKA,[29] SVM - support vector machines as implemented in LibSVM.[36] Dragon - descriptor software by Talete inc.,[25] SMF - substructural molecular fragments as implemented in ISIDA,[21] E-state - electrotopological state indices.[24]

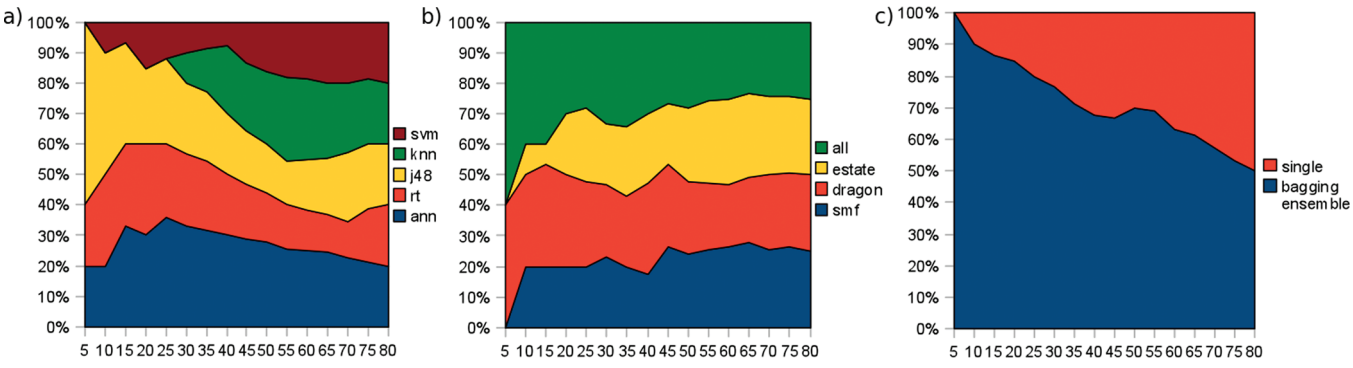


Figure 1. cumulative charts of share of models of each type among the top-performing models. The horizontal axis displays the amount of top performing models taken into account; the vertical axis displays a share of each type of machine learning methods, descriptors or ensemble approaches among these models. Larger areas (J48 and ANN, Dragon and All, Bagging) demonstrate more successful approaches.

Comparison of machine learning methods. Among the used methods, best performances were achieved with J48, ASNN and RT methods. J48 and ASNN produced the models with the highest performance using the full set of descriptors. However, when considering the applicability domain of models and 90% threshold, as discussed below, the ASNN method performed better. There were also several other models that were statistically nonsignificantly different to the most accurate ones with the significance level of 0.05. The models produced by the SVM and KNN methods were significantly less accurate as compared to the top-performing ones. The KNN was also the only machine learning method that didn't produce a model within top 20 most accurate models. Thus, this method had a lower performance than other machine learning methods analyzed in our work.

Comparison of descriptors. The full combined set of descriptors showed the highest accuracy in most cases. When used separately, Dragon descriptors demonstrated the highest performance, with E-State and SMF performing approximately equally accurate. These results show, that a combination of the descriptor sets calculated with different approaches brings new information to the model and increases its performance. It is important to note that Dragon (and, as a result, the full set)

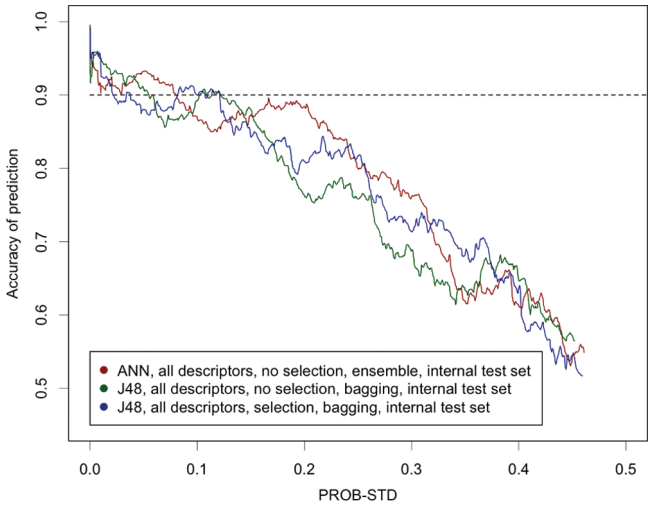


Figure 2. Prediction accuracy of a model for a particular group of molecules from an internal test set as a function of PROB-STD. Different colors stand for different models.

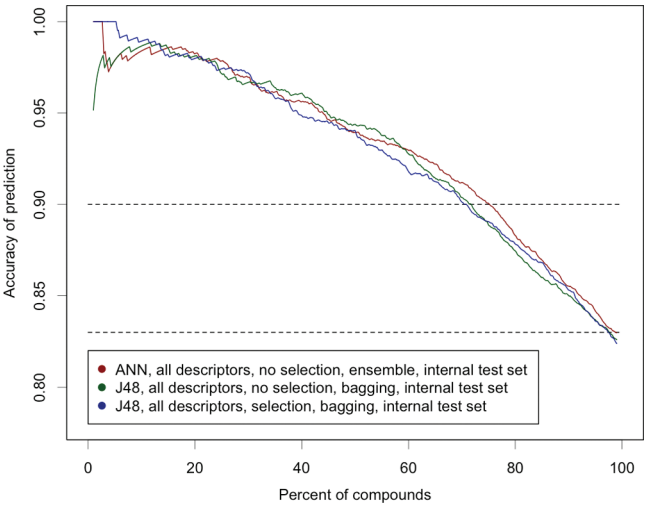


Figure 3. Average prediction accuracy of a model on a fraction of the internal test set as a function of the size of this fraction (compounds ordered by PROB-STD). Different colors stand for different models.

includes 3D descriptors. Therefore the high performance values for models including Dragon descriptors may demonstrate the importance of 3D information for modeling CYP1A2 inhibition activity. On the other hand, the generation of 3D structures can be a limiting step and can significantly increase the time required for application of models using these sets of descriptors.

Bagging/ensembles work better compared to the single models. The charts at Figure 1 demonstrate that bagging/ensemble methods performed better then single-models. Table 1 confirms this result and also indicates that bagging and ensemble approaches significantly improved the performance of ASNN, RT and J48 models. However, these approaches had less or no influence on the KNN and SVM models. The KNN and SVM methods are more stable and are less influenced by distortions of the training set due to bagging. The former three methods, however, have a stronger intrinsic variability and models calculated with such methods using different bagging replica have larger variations. The standard deviations of predictions for test set molecules were 0.37, 0.34, and 0.32 for RT, ASNN and J48 methods, respectively, while they were only 0.12 and 0.15 for KNN and SVM, respectively. These standard deviations were calculated using 100 models from ensemble (ASNN) and bagging (all other methods), respectively. This result indicates that methods with higher variation of predictions (RT, ASNN and J48) had a higher gain from using ensemble approach, as it is clear from Table 1. This result is in agreement of previous conclusions of Breiman,[37] who reported similar results by considering bias and variance of models. He assumed that methods with higher variation of results may have lower bias and their low performance could be mainly due to higher variation of their predictions. The average of predictions of such methods decreases their variance and improves their accuracy. The performances of more stable methods (e.g., SVM, KNN) are to a larger degree dependent on their biases. Therefore, the use of ensemble approach does not improve their accuracy.

The increase of model accuracy came at a price of increasing usage of computational resources both for training and application of a model. In the presented study the bag (for RT and J48) and an ensemble (for ASNN) consisted of 100 model instances. This led to a 100 times increase of computational time required to create and apply these models.

Descriptor selection does not improve results. In this study the models trained on the preselected set of descriptors performed generally worse, than those trained on the full set of descriptors. These results indicate that the used descriptor selection strategy was not the optimal for the analyzed methods. These results contradict the conclusions of another study[16] where the same strategy was found to increase the accuracy of models. However, at the same time we found that the descriptor selection at least did not decrease the performance of RT and J48 methods, which performed nonsignificantly different with or without variable selection for some sets of descriptors.

Applicability domain of models. Figures 2 and 3 show two different chart types that illustrate the ability to differentiate accurate and inaccurate predictions for CYP1A2 models using the PROB-STD DM.

The charts are plotted for the internal test set compounds sorted by PROB-STD. Figure 2 displays the accuracy of predictions calculated as simple moving average over a window of 200 compounds. The plot shows the percentage of correct predictions in a window for each particular value of PROB-STD measure. The plot has a general downward trend that shows a strong correlation of the prediction accuracy and a DM.

Figure 3 represents cumulative accuracy-coverage plots. This chart displays prediction accuracy (y axis) for a group of compounds, having DM less than some threshold against percentage of this group of compounds in the whole set (x axis). The plot starts from high accuracy values (for compounds with low PROB-STD measures) and drops to the level of 0.83 - the average accuracy for the whole set. In particular, this chart shows us, that top 70%-75% of the internal test set, ordered by PROB-STD values, can be predicted with an average accuracy of 90%.

As we can see, the behavior of the plots is similar for different models. This means the PROB-STD DM worked universally, and was successfully used with all sets of descriptors and machine learning methods, as long as ensemble or bagging approach was taken.

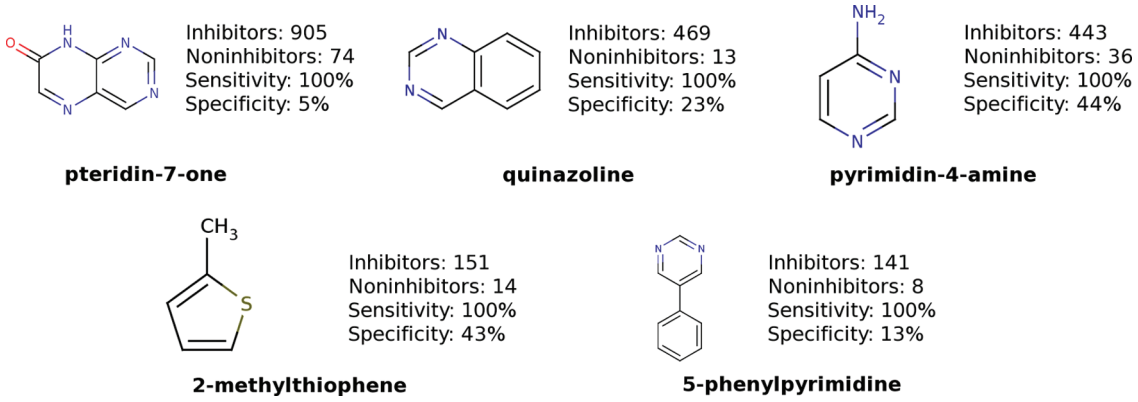


Figure 4. Inhibitor-specific fragments of the data set.

Interpretation of modeling results. One more common area of application of QSAR models, except for early stage filtering of potential drug candidates, is guiding the drug candidate optimization process. Easily interpretable models may provide useful insight on the structural changes to the molecule that are required for it to obtain some desired property. In this study we adopted a fragment-based interpretation approach, as described in the "Methods" section.

Figure 4 displays the inhibitor-specific fragments that appear in the biggest amount of molecules in a data set along with sensitivity/specificity values. The sensitivity and specificity values are given for the whole data set of 7486 molecules, model results are obtained from a 5-fold cross-validated model built by the "best approach" - ASNN, the full descriptor set with no descriptor selection. That is, the initial full data set was randomly split to five folds and five ASNN models were built, one of the folds participating as a test set and the other four combined — a training set for each single model. This way we receive valid predictions for the whole data set, i.e. each model in 5-fold cross-validation predicted 20% molecules.

The fragments at Figure 4 account for 2123 inhibitors, which is around 53% of all inhibitors in the data set. All these molecules were correctly classified as inhibitors. This number is smaller than the sum of individual fragment values since some molecules contain two or three of the presented fragments.

A total number of 145 noninhibitor molecules contained these five fragments. Only 30 of them were correctly classified as noninhibitors.

The common pattern for these fragments is 100% sensitivity (all inhibitor molecules containing these fragments were correctly classified as inhibitors) but low 5 — 40% specificity (only 30 out of 145 noninhibitor molecules containing these fragments were correctly classified as noninhibitors, which results to overall 21% specificity for the molecules containing these fragments).

The fragments in Figure 4 are clearly associated with inhibition activity. This result is in agreement with other studies of CYP1A2, which indicate that planar aromatic groups participating in $\pi-\pi$ interactions are the essential requirement for CYP1A2 substrates and inhibitors.[7,42−45] It is interesting that four of five identified fragments contain pyrimidine fragment. This finding provides a testable hypothesis on the importance of these fragments for CYP1A2 inhibition that can be verified experimentally.

External set result. In this part of the study we applied the models to an external data set. As a test set we used 160 inhibitors of CYP1A2 obtained from literature. These data were measured using different approaches and etalon reactions and thus were expected to have a high amount of variability. The molecules from this data set were diverse and did not appear in the training set for the evaluated models. The overall accuracy of prediction for this external data set was 59% - 68% of correctly classified

Table 2. The performance of best 20 models for the external test set of 160 of CYP1A2 inhibitors

	DESCR	SEL	METHOD	ENSEMBLE	ACC	ACC (AD)
1	All	no	ASNN	ensemble	0.68	*0.90*
2	All	no	J48	bagging	0.65	*0.87*
3	All	yes	J48	bagging	0.66	*0.85*
4	Dragon	no	J48	bagging	0.65	*0.82*
5	Dragon	yes	RT	bagging	0.70	*0.85*
6	All	yes	RT	bagging	0.68	*0.80*
7	SMF	no	J48	bagging	0.66	*0.87*
8	Dragon	no	ASNN	ensemble	0.68	*0.80*
9	SMF	no	SVM	single	0.65	-
10	E-state	no	RT	bagging	0.63	*0.82*
11	Dragon	yes	J48	bagging	0.61	*0.77*
12	All	yes	ASNN	ensemble	0.62	*0.74*
13	SMF	no	ASNN	ensemble	0.63	*0.77*
14	Dragon	no	ASNN	single	0.61	-
15	All	no	RT	bagging	0.59	*0.77*
16	E-state	no	ASNN	ensemble	0.69	*0.85*
17	E-state	no	SVM	bagging	0.61	*0.69*
18	E-state	no	SVM	single	0.60	-
19	Dragon	no	RT	bagging	0.61	*0.69*
20	SMF	no	RT	bagging	0.65	*0.77*

ASNN — Associative Neural Networks, RT and J48 — random trees and C4.5 pruned trees as implemented in WEKA,[29] SVM - support vector machines as implemented in LibSVM.[36] Dragon - descriptor software by Talete inc.,[25] SMF - substructural molecular fragments as implemented in ISIDA,[20] E-state - Electrotopological state indices.[24] ACC — average model accuracy on an external test set. ACC(AD) — average model accuracy on a part of an external test set within AD of the model.

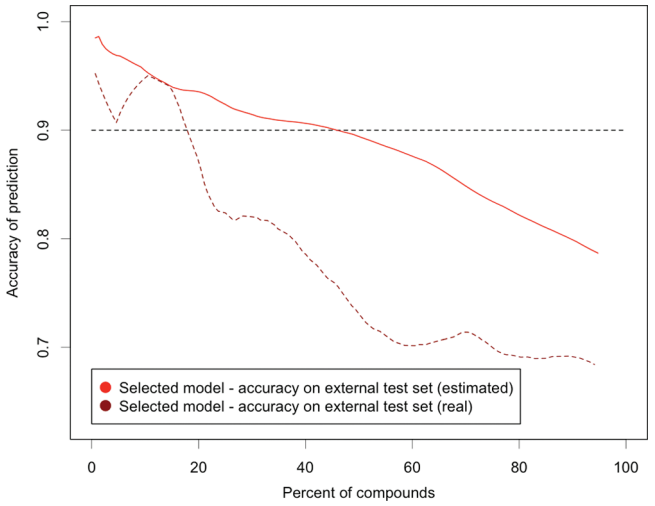


Figure 5. Average estimated and calculated accuracies of the ASNN model (Table 1) for the external test sets for compound predictions ordered by PROB-STD.

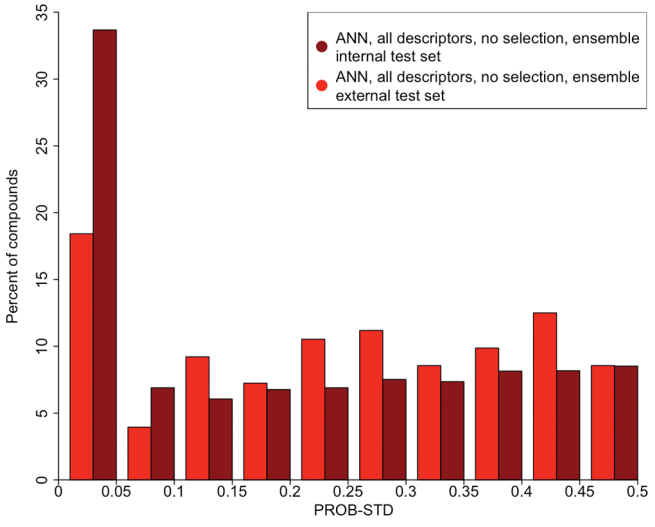


Figure 6. Distribution of number of compounds by PROB-STD values.

instances. The WACC and MCC measures could not be calculated because the set contained only instances of one class (inhibitors).

Table 2 shows the external set accuracy of the 20 top models. It also displays the accuracy of those molecules, which where in the applicability domain of the models (as described in the AD-related section of the article).

Figure 5 represents cumulative accuracy-coverage plot of expected and observed accuracies of the ASNN model built on full descriptor set with no descriptor selection.

We can see in Figure 5 that on the qualitative level the behavior of the accuracy-coverage plot of the external test set is similar to that of the internal test set. Apparently the accuracy of the model on the external test set was significantly lower than on the internal test set. This is not surprising, since the diversity of molecules in this set is higher. If we order the compounds by PROB-STD, the accuracy increases to 90% of correctly classified compounds for about 20% external test set molecules. When we take into account the whole 100% of compounds, the accuracy drops to 68%.

Figure 6 shows the distribution of molecules in the internal and external test sets by PROB-STD values. Both sets have a high amount of molecules (18% for the external test set and 34% for the internal one) with PROB-STD between 0 and 0.05 - these molecules contribute to the highest accuracy of predictions. For the external test set, the molecules are distributed more evenly over the range of PROB-STD values - a consequence of a lower overall accuracy of the model on this set.

CONCLUSIONS

In this paper, different QSAR approaches for the prediction of CYP1A2 inhibition were compared. Dragon, E-State and ISIDA SMF descriptors were used. The kNN, SVM, ASNN, RT and J48 methods were studied. Models built on part of PubChem BioAssay data set were applied to predict a test set from the same source, as well as an external data set, collected from literature.

ASNN neural networks in combination with the full descriptor set calculated the highest accuracy, which was 83% of correctly classified instances on the internal test set. This result is about 5% higher than the previously published results using the same data set.[16] Several other models (including J48 and RT in combination with bagging approach) showed results, statistically nonsignificantly different to the top performance model.

In a majority of cases the models built on the full set of descriptors outperformed the models built on preselected sets of descriptors.

For neural networks, random tree and J48 tree the bagging/ensemble approach allowed a statistically significant increase of performance.

Increasing the number of different types of descriptors had the positive effect on the model accuracy. The most accurate models in this study included Dragon 3D descriptors. This showed the importance of molecule 3D information for the modeling of CYP1A2 inhibition activity.

The models were applied to the heterogeneous external test set and achieved the accuracies of 59% - 68% correctly classified instances. This result is expected since the molecules in external set were less similar to the training set compounds compared to the molecules used as the internal test set according to PROB-STD DM (see Figure 6). The lower accuracy of the model for the external set can be also explained by a wide variety of protocols and criteria used to measure CYP1A2 inhibition activity. Using the PROB-STD measure allowed us to identify about 20% of external set compounds, for which the average accuracy of predictions was 90%.

Several molecular fragments in the studied data set (such as pteridin-7-one, quinazoline, 2-methyltiophene) were mostly present in CYP inhibitors. This result is in correspondence with numerous CYP1A2 studies using docking approaches, which indicate an importance of $\pi-\pi$ interactions for CYP1A2 inhibition.

The models were created using the Online Chemical Modeling Environment (OCHEM).[41] The top performing models are available at OCHEM at http://ochem.eu/models/Q5747.

AUTHOR INFORMATION

Corresponding Author

*Dr Igor V. Tetko, Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Ingolstädter

Landstrasse 1, D-85764 Neuherberg, Germany itetko@vcclab.org Tel.: +49-89-3187-3575 Fax: +49-89-3187-3585.

■ ACKNOWLEDGMENT

This study was partially supported with GO-Bio BMBF grant 0313883, FP7 project CADASTER 212668, and Germany-Ukraine collaboration project UKR 08/006. The authors would like to thank Volodymyr Prokopenko, Vasyl Kovalsihyn and Larisa Charochkina from Germany-Ukraine collaboration project UKR 08/006 for collecting some of experimental data used as the external test set.

■ REFERENCES

(1) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

(2) Masimirembwa, C. M.; Thompson, R.; Andersson, T. B. In vitro high throughput screening of compounds for favorable metabolic properties in drug discovery. *Comb. Chem. High T. Scr.* **2001**, *4*, 245–263.

(3) Pelkonen, O.; Turpeinen, M.; Hakkola, J.; Honkakoski, P.; Hukkanen, J.; Raunio, H. Inhibition and induction of human cytochrome P450 enzymes: current status. *Arch. Toxicol.* **2008**, *82*, 667–715.

(4) Guengerich, F. P.; Wu, Z.-L.; Bartleson, C. J. Function of human cytochrome P450s: Characterization of the orphans. *Biochem. Bioph. Res. Co.* **2005**, *338*, 465–469.

(5) Pirmohamed, M.; Park, B. K. Cytochrome P450 enzyme polymorphisms and adverse drug reactions. *Toxicology* **2003**, *192*, 23–32.

(6) Michalets, E. L. Update: clinically significant cytochrome P-450 drug interactions. *Pharmacotherapy* **1998**, *18*, 84–112.

(7) Lewis, D. F. 57 varieties: the human cytochromes P450. *Pharmacogenomics* **2004**, *5*, 305–318.

(8) Wolf, C. R.; Smith, G. Pharmacogenetics. *Brit. Med. Bull.* **1999**, *55*, 366–386.

(9) Pelkonen, O.; Mäenpää, J.; Taavitsainen, P.; Rautio, A.; Raunio, H. Inhibition and induction of human cytochrome P450 (CYP) enzymes. *Xenobiotica* **1998**, *28*, 1203–1253.

(10) Flockhart D. A. *Drug Interactions: Cytochrome P450 Drug Interaction Table. Indiana University School of Medicine*. http://medicine.iupui.edu/clinpharm/ddis/table.asp (accessed Apr 26, 2011).

(11) Wang, B.; Zhou, S.-F. Synthetic and natural compounds that interact with human cytochrome P450 1A2 and implications in drug development. *Curr. Med. Chem* **2009**, *16*, 4066–4218.

(12) Burton, J.; Ijjaali, I.; Barberan, O.; Petitet, F.; Vercauteren, D. P.; Michel, A. Recursive partitioning for the prediction of cytochromes P450 2D6 and 1A2 inhibition: importance of the quality of the dataset. *J. Med. Chem.* **2006**, *49*, 6231–6240.

(13) Gleeson, M. P.; Davis, A. M.; Chohan, K. K.; Paine, S. W.; Boyer, S.; Gavaghan, C. L.; Arnby, C. H.; Kankkonen, C.; Albertson, N. Generation of in-silico cytochrome P450 1A2, 2C9, 2C19, 2D6, and 3A4 inhibition QSAR models. *J. Comput. Aided Mol. Des.* **2007**, *21*, 559–573.

(14) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome p450 substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588–2605.

(15) Dagliyan, O.; Kavakli, I. H.; Turkay, M. Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 2403–2411.

(16) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P. E.; Olsen, L.; Jørgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37*, 658–664.

(17) 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.

(18) National Library of MedicineNational Institute of Health *PubChem BioAssay AID-410*. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=410 (accessed Apr 26, 2011).

(19) Promega *P450-Glo(TM) Assays*. http://www.promega.com/tbs/tb325/tb325.html (accessed Apr 26, 2011).

(20) Chemaxon *Standardizer, JChem 5.4*. http://www.chemaxon.com (accessed Apr 26, 2011).

(21) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solovev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - Platform for Virtual Screening Based on Fragment and Pharmacophoric Descriptors. *Curr. Comput.-Aid. Drug.* **2008**, *4*, 191–198(8).

(22) Solov'ev, V. P.; Varnek, A.; Wipff, G. Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847–858.

(23) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.

(24) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press, 1999, pp 1−239.

(25) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics: Vol. I: Alphabetical Listing/Vol. II: Appendices, References*; 2nd ed.; Wiley-VCH, 2009, p 232.

(26) Molecular Networks GmbH: ErlangenGermany *CORINA* http://www.molecular-networks.com/ (accessed Apr 26, 2011).

(27) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000–1008.

(28) Mishra, N.; Agarwal, S.; Raghava, G. Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule. *BMC Pharmacol.* **2010**, *10*, 8.

(29) University of Waikato: WaikatoNew Zeland *Weka: Waikato Environment for Knowledge Analysis*. http://www.cs.waikato.ac.nz/ml/weka/ (accessed Apr 26, 2011).

(30) Hall, M. A. Correlation-based Feature Subset Selection for Machine Learning, University of Waikato: Hamilton, New Zealand, 1998.

(31) Tetko, I. V. Associative neural network. *Methods Mol. Biol.* **2008**, *458*, 185–202.

(32) Tetko, I. V. Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717–728.

(33) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman & Hall, New York, NY, 1984, pp 1−171.

(34) Quinlan, R.; Quinlan, J. R. *C4.5: Programs for Machine Learning*; Revised, Update.; Morgan Kaufman Publ Inc, 1993, pp 1−109.

(35) Vapnik, V. N. *Statistical Learning Theory*; Wiley-Interscience, 1998, pp 1−736.

(36) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2001.

(37) Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140.

(38) Good, P. I. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*; 3rd ed.; Springer, 2004, pp 1−276.

(39) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J. Chemometr.* **2010**, *24*, 202–208.

(40) Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Cherkasov, A.; Li, J.; Gramatica, P.; Hansen, K.; Schroeter, T.; Müller, K.-R.; Xi, L.; Liu, H.; Yao, X.; Oberg, T.; Hormozdiari, F.; Dao, P.; Sahinalp, C.; Todeschini, R.; Polishchuk, P.; Artemenko, A.; Kuz'min, V.; Martin, T. M.; Young, D. M.; Fourches, D.; Muratov, E.; Tropsha, A.; Baskin, I.; Horvath, D.; Marcou, G.; Muller, C.; Varnek, A.; Prokopenko, V. V.; Tetko, I. V. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J. Chem. Inf. Model* **2010**, *50*, 2094–2111.

(41) Sushko et al. Online Chemical Modeling Environment (OCHEM): Web Platform for Data Storage, Model Development and Publishing of Chemical Information (in press) *J. Comput. Aided Mol. Des.*

(42) 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.

(43) Sansen, S.; Yano, J. K.; Reynald, R. L.; Schoch, G. A.; Griffin, K. J.; Stout, C. D.; Johnson, E. F. Adaptations for the Oxidation of Polycyclic Aromatic Hydrocarbons Exhibited by the Structure of Human P450 1A2. *J. Biol. Chem.* **2007**, *282*, 14348–14355.

(44) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates Part 1: active site characteristics. *Drug Discov. Today* **1997**, *2*, 406–414.

(45) Smith, D. A.; Ackland, M. J.; Jones, B. C. Properties of cytochrome P450 isoenzymes and their substrates part 2: properties of cytochrome P450 substrates. *Drug Discov. Today* **1997**, *2*, 479–486.