Classification of Cytochrome P450 1A2 Inhibitors and Non-Inhibitors by Machine Learning Techniques

Poongavanam Vasanthanathan, Olivier Taboureau, Chris Oostenbrink, Nico P. E. Vermeulen, Lars Olsen and Flemming Steen Jørgensen

Biostructural Research, Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark (P.V., L.O., F.S.J.); Leiden-Amsterdam Center for Drug Research, Section of Molecular Toxicology, Department of Chemistry and Pharmacochemistry, Vrije Universiteit, De Boelelaan 1083, 1081 HV, Amsterdam, The Netherlands (P.V., C.O., N.P.E.V.) and Chemoinformatics, BioCentrum-DTU, Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark (O.T.).
Running title: Classification of Cytochrome P450 1A2 Inhibitors and Non-Inhibitors

Address correspondence to: Flemming Steen Jørgensen, Biostructural Research, Department of Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Copenhagen, Universitetsparken 2, DK-2100 Copenhagen, Denmark Tel.: +45 3533 6378, Fax: +45 35 33 60 01, E-mail: fsj@farma.ku.dk

Manuscript statistics:

- Number of text pages: 24
- Number of tables: 6
- Number of figures: 6
- Number of references: 26
- Number of words in Abstract: 210
- Number of words in Introduction: 450
- Number of words in Results and Discussion: 1725

Abbreviations: CYP1A2: cytochrome P450 1A2 isoform; QSAR: quantitative structure activity relationship; DOOD: D-optimal onion design; kNN: kappa nearest neighbors; SVM: support vector machine; DT: decision tree; RF: Random forest; MOE: molecular operating environment; HTS: high throughput screening; 2D/3D: two and three dimension respectively; PCA: principle component analysis; $R^2$: correlation coefficient; $Q^2$: cross validated correlation coefficient; MCC: Mathews correlation coefficient.
Abstract

The cytochrome P450 (CYP) superfamily plays an important role in the metabolism of drug compounds, and it is therefore highly desirable to have models that can predict whether a compound interacts with a specific isoform of the CYPs. In this work, we provide in silico models for classification of CYP1A2 inhibitors and non-inhibitors. Training and test sets consisted of about 400 and 7000 compounds, respectively. Various machine learning techniques, like binary QSAR, support vector machine (SVM), random forest, kappa nearest neighbors (kNN), and decision tree methods were used to develop in silico models, based on Volsurf and MOE descriptors. The best models were obtained using the SVM, random forest, and kNN methods in combination with the BestFirst variable selection method, resulting in models with 73 – 76 % of accuracy on the test set prediction (Matthews Correlation Coefficient of 0.51 and 0.52). Finally, a decision tree model based on Lipinski’s Rule-of-five descriptors was also developed. This model predicts 67 % of the compounds correctly and gives a simple and interesting insight into the issue of classification. All the developed models in this work are fast and precise enough to be applicable for virtual screening of CYP1A2 inhibitors or non-inhibitors, or can be used as simple filters in the drug discovery process.
Introduction

Cytochromes P450 (CYPs) are heme-containing enzymes found in both prokaryotes and eukaryotes, and they are involved in a wide range of cellular biotransformation functions. From a pharmaceutical perspective, the most important function is the degradation of drugs (Nebert et al., 2002). In general, hydrophobic compounds are converted into more hydrophilic species, to facilitate excretion.

The most important CYP isoforms being involved in metabolism of drugs in humans are CYP1A2, CYP2A6, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4. CYP1A2 constitutes 12 % of the total CYP content in the liver and plays an important role in the metabolic clearance of ~5 % of currently marketed drugs. The substrates for the CYP1A subfamily are generally characterized as neutral, flat, aromatic and lipophilic (two to four aromatic rings) with at least one putative hydrogen bond donor (Smith et al., 1997), in agreement with the observed contacts in the recent crystal structure of CYP1A2 (Sansen et al., 2007). Examples of drugs being CYP1A2 substrates are acetaminophen, caffeine, cloazepine, haloperidol, olanzapine, propranolol, tacrine, theophylline, and zolmitriptan (Flockhart, 2007).

In silico approaches are attractive as they can be used in an early stage of the drug discovery process and thereby reduce the number of experimental studies and improve the success rates. For this purpose, various traditional in silico modeling methods and more recently developed non-linear machine learning methods have been used (Chohan et al., 2005; Eitrich et al., 2007; Fox and Kriegl, 2006; de Graaf et al., 2005; Kriegl et al., 2005a; Terfloth et al., 2007; Yap and Chen, 2005; Yap et al., 2006; Zhou et al., 2007). Machine learning methods are particularly useful for data mining of large databases to discover patterns or rules to derive models for problems where the underlying mechanism is not clear. For example, support vector machine (SVM) methods have
been applied to classify inhibitors of the CYP3A4 enzyme with a success rate of about 70% for the test set (807/538 compounds in training/test sets) (Kriegl et al., 2005a) and to predict isoform specificity of CYP3A4, 2D6, and 2C9 substrates with about 80% of the test set correctly predicted (146/233 compounds in training/test sets) (Terfloth et al., 2007).

Thus, the *in silico* methods seem promising for making reliable models for sets of a large number of compounds. In this study, we use about 400 compounds to construct models for CYP1A2 inhibition, and explore the accuracy of various machine learning methods like support vector machine (SVM), random forest, decision tree, kappa nearest neighbors (kNN), and binary QSAR methods for a test set containing 7000 compounds. This is to our knowledge the first time that the information from such a large number of compounds is used to generate and validate *in silico* models for classification of CYP1A2 inhibitors and non-inhibitors.
Materials and Methods

Biological activity. The structures and biological activity of 8,342 compounds were collected from the PubChem bioassay database (http://www.ncbi.nlm.nih.gov, accessed July 2007). In brief, the inhibition of human CYP1A2 catalyzing the demethylation of luciferin 6'-methyl ether to luciferin was measured. The luciferin formation is monitored by luminescence after the addition of a luciferase detection reagent. The luciferin 6'-methyl ether concentration in the assay was equal to its $K_M$ for CYP1A2. Out of the 8,342 compounds downloaded from PubChem, 4,173 were assigned as active (inhibitors), 3,514 as inactives (non-inhibitors), and 655 as inconclusive with respect to the activity. The inconclusives were not considered in this study to avoid uncertainty in the model building. In addition to the PubChem data set, we also used an external test set of 89 drug compounds, kindly provided to us by Vertex Pharmaceuticals for validation of the developed models (Zlokarnik et al., 2006).

3D structure generation. The 2D structures were converted into 3D structures by using the CONCORD software (version 6.1.2, accessed via SYBYL from Tripos Associates Inc., St. Louis, USA). Of the 7,687 structures, 218 could not be converted to 3D due to restrictions in the CONCORD program (number of rotatable bonds > 25, ring size > 10, number of atoms > 80 or the presence of atoms like Hg, As, Sb etc.) or incorrect or insufficient stereochemical information in the 2D notation from PubChem. For the remaining 7,469 structures various descriptors were calculated.

MOE 2D descriptors. All 3D structures were imported into the molecular modeling software Molecular Operating Environment (MOE) (version 2006.08, Chemical Computing Group Inc, Montreal, Quebec, Canada). They were pre-processed by removing all counter ions, solvent
molecules, and salt in the structures. Subsequently, the structures were geometry optimized using the MMFF94s force field and 214 2D descriptors implemented in MOE were computed (representing atom and bond counts, Kier and Hall connectivity, kappa shape indices, adjacency, distance matrix descriptors, pharmacophore feature descriptors, partial charge descriptors, potential energy descriptors, surface area, volume, and shape descriptors) (Labute, 2000).

**Volsurf descriptors.** The geometry optimized compounds were imported into the Volsurf software (version 4.1.4.1, Molecular Discovery Ltd., Middlesex, United Kingdom). Volsurf is a computational procedure to transform 3D molecular interaction fields from a GRID calculation into quantitative numerical descriptors. For the present study, we used the following probes: water (H$_2$O), hydrophobic (DRY), H-bonding carbonyl (O), and N1 amide (N).

**Selection of training and test set.** All 214 2D MOE descriptors were merged with the 110 Volsurf descriptors for the total set of 7,469 compounds. The total data set was divided into a training and a test set. The training set was selected by use of D-optimal onion design (DOOD). DOOD is a method for selecting a training set of reasonable size which is representative for the chemical property space defined by the molecular structures (Kriegel et al., 2005b; Olsson et al., 2004). In brief, the general idea of DOOD is to score vectors calculated by either principal component analysis (PCA) or partial least squares (PLS) and then to split the dataset into different layers (or subsets) enabling the selection of training set compounds from each layer. This ensures that representative molecules are selected from the inner layers of the onion design and that the training set contains a diverse range of chemical structures in addition to covering the range of chemical properties. The total number of compounds to be selected can be controlled by the number of layers and the type of regression model targeted within each layer. The performance of DOOD has already been assessed with success in drug discovery. Compared to a random selection, which may...
introduce bias and variability if the selected training set is not balanced, the DOOD method has shown a better stability for the prediction models (Olsson et al., 2004).

In this work, we used the scores from the principal component analysis (PCA) obtained by the SIMCA-P program (version 11.0, Umetrics, Umeå, Sweden) ($R^2=0.81$, $Q^2=0.66$, 25 principal components (PCs)). The layers were determined according to the Hotelling’s $T^2$ parameter which measures the distance between the projection of a compound to the center of the model. The compounds were then allocated into the individual layers according to their Hotelling’s $T^2$ values for the corresponding model. The software package MODDE (version 7.0, Umetrics, Umeå, Sweden) was used to perform the DOOD-based training set selection.

We also divided the full set of compounds into a training and a test set using another approach (the kNN methods, see below for a description of the method). The quality of the final models were almost identical to the models developed with the training set selected with DOOD (data not shown).

**Machine learning methods.** The selected methods are based on different concepts and representative of the common approaches considered for classification. The machine learning methods applied in this work are briefly presented below.

**Support vector machine (SVM).** SVM is a non-linear model, developed by Vapnik (Vapnik, 1995). The SVM method constructs a hyperplane which discriminates between data points of distinct classes (binary SVM) such that the margin between both classes is maximized. This margin models the linear decision hyperplane. The final position and orientation of the hyperplane is defined by a subset of training vectors, the so-called support vectors. The support vector machine (SVM) approaches were used in association with a radial bias function as the kernel function. The
kernel exponent was set to 1.0 (SVM^D-Default) and the polynomial kernel to 2.0 (SVM^E) for linear and non-linear SVM classifiers, respectively.

*Kappa nearest neighbor algorithm (kNN).* This method was developed by Zheng and Tropsha (2000). The k-nearest neighbor models are based on the assumption that similar compounds have similar physico-chemical and biological activity profiles and can define a class membership of its nearest neighbors. Compounds are represented by their position vectors defined in the physico-chemical space and then biological activity is assigned by a majority vote of its neighbors. The neighbors are taken from a set of compounds for which the correct classification is known. Finally, properties or activities of new compounds are assigned based on the majority vote of neighbors defined previously. For the kappa nearest neighbor algorithm, we used Euclidian distances and 5 kappa nearest neighbors to avoid tied votes.

*Random forest method.* With the random forest method (Witten and Frank, 2005), multiple classification trees are constructed from an input vector. The input vector is then placed down of the classification trees in the forest and each tree gives a classification, or votes, for that class. Finally, the random forest chooses the classification having the most votes. The forest error rate depends on two things: 1) the correlation between any two trees in the forest 2) the strength of each individual tree in the forest. A tree with a low error rate is a strong classifier. The main advantage of this approach is the possibility to handle thousands of input variables without overfitting in a very fast way. The random forest method (RF) was used with ten trees and one seed.

*Decision tree (C4.5/J48).* This method is a divide-and-conquer approach to the problem of learning from a set of independent instances that naturally leads to a style of representation. First, an
attribute is selected to place at the root node and make one branch for each possible value. This splits up the data set into subsets, one for every value of the attributes. This process can be repeated recursively for each branch, using only those instances that actually reach the branch (Witten and Frank, 2005).

**Binary QSAR.** This method is based on Bayesian inference technique, which is used to classify whether a compound is active or inactive on the basis of its associated molecular descriptors. A probabilistic distribution of active and inactive compounds in a training set is determined using a partial least squares (PLS) method. The derived binary QSAR model can subsequently be used to predict the probability of new compounds to be active against given targets. A detailed description of the binary QSAR methodology has been described in the literature (Gao et al., 1999; Gao, 2001; Labute, 1999).

The influence of various factors on the accuracy of the predictions was investigated. A smoothing factor was used to minimize the sensitivity of the derived model to the selection of binary boundaries. We analyzed different smoothing factors (Gao et al., 1999; Gao, 2001; Labute, 1999) ranging from 0.08 to 0.25 with different numbers of principal components ranging from 1 to maximum (data not shown). For descriptor selection, different variable combinations were evaluated. Finally, we selected 110 (MOE/Volsurf) descriptors for model development. The descriptors include MOE 2D descriptors like subdivided surface area descriptors (SlogP descriptors), number of aromatic atoms, number of hydrogen atoms, atomic valence connectivity index, carbon valence connectivity index, molecular weight and Balaban’s connectivity topological index and Volsurf descriptors using the DRY and N1 probes.
Software. The Weka data mining software (version 3.2, Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/ml/Weka/) (Witten and Frank, 2005) was used for the inhibitor/non-inhibitor classification. The software provides a set of classification and regression methods, variable selection methods, and clustering methods (SVM\(^D\), SVM\(^E\), kNN, random forest, and decision tree). The binary QSAR (BQ) was performed using the MOE programme.

Attribute selection. For attribute selection we used the automatic variable selection procedure (CfsSubsetEval) in the Weka software. CfsSubsetEval was combined with either the BestFirst or a genetic algorithm (Witten and Frank, 2005).

Matthews correlation coefficient (MCC). MCC is a measure of the quality of a binary classification. It takes into account true positives and negatives and is generally regarded as a balanced measure which can also be used if the classes are of very different sizes. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction, 0 an average random prediction and -1 the worst possible prediction. Generally, MCC values above 0.4 are considered to be predictive in machine learning methods (Chohan et al., 2005).

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

where \( TP \) are true positives, \( TN \) are true negatives, \( FP \) are false positives, and \( FN \) are false negatives.

Cross validation. All developed models were cross validated with \( K \)-fold cross-validation. In \( K \) fold cross validation, the original sample is divided into \( K \) subsamples. Of the \( K \) subsamples, a single subsample is retained as the validation data for testing the model, and the remaining \( K - 1 \)
subsamples are used as training data. The cross-validation process is then repeated $K$ times, with each of the $K$ subsamples used exactly once as the validation data. Here, we performed fivefold cross validation. All cross validations were carried out with the procedures in the Weka software, and 5-fold cross validation was chosen.
Results and Discussion

Description of training and test sets

The 7469 compounds extracted from PubChem and considered in this work constitute the largest set of compounds used for constructing in silico models of CYP1A2. Visual inspection of randomly selected compounds revealed compounds resembling drug compounds as well as compounds being clearly non-drug-like. Although Lipinski's Rule-of-five (Lipinski et al., 2001) is not a direct measure of drug-likeness, we calculated the violations of each of the four rules for the CYP1A2 dataset: 544 compounds had molecular weight larger than 500; 132 compounds had more than 10 hydrogen bond acceptors; 293 compounds had more than 5 hydrogen bond donors and 551 compounds had a partition coefficient, logP value, above 5. In total, only about 15% of the compounds in the dataset violated one or several of the Lipinski's rules.

From D-optimal onion design (DOOD), a training set consisting of 411 compounds (192 inhibitors and 219 non-inhibitors) was constructed, and the remaining 7,058 compounds were used as a test set. DOOD was used because it has previously been successfully applied for the selection of suitable training sets (Gavaghan et al. 2007). The score plot from the PCA shows that the diversity of the whole dataset is satisfactory reflected in the training set (cf. Figure 1). Only a minority of the compounds are outside the 95% confidence interval, and of those, most form a cluster located in the lower right part of the score plot, corresponding to the compounds violating Lipinski’s Rule-of-Five. Inspection of these structures revealed that there is no structural similarity within this group, but that 95% of these compounds were non-inhibitors.

Construction of binary QSAR models

Binary QSAR models for classification of active (inhibitors) and inactive (non-inhibitors) compounds were developed using the 214 calculated MOE and 110 Volsurf descriptors. We
selected these descriptors because they previously have been applied successfully for classification or IC₅₀ predictions for CYP ligands (Kriegl et al. 2005b, Eriksson et al. 2004). Most of these descriptors cover properties like lipophilicity (e.g. LogP-derived descriptors or Volsurf descriptors based on the hydrophobic interactions – the so-called DRY probe), the hydrophilicity (e.g. polar surface area derived descriptors or Volsurf descriptors based on the hydrophilic interactions – the so-called N1 probe), or the number of various types of atoms (e.g. number of aromatic atoms, hydrogen atoms).

The influence of various factors like the number of descriptors, the number of principle components (PCA), and smoothing factors on the accuracy of the predictions was investigated (cf. Materials and Methods section). Finally, we selected 110 (MOE and Volsurf) descriptors for model development, including those describing the lipophilicity, hydrophilicity, various atom counts, connectivity, etc. Using default settings in MOE for smoothing factors and number of PCs gave a binary QSAR model with a significant difference in predictive power between the training and test set, in particular for the CYP1A2 inhibitors, for which 96 % are correctly predicted in the training set, but only 55 % in the test set. Therefore, a model that is based on only three principal components was considered. This leads to a correctly predicted number of compounds on 70 % and 66 % for the training and test sets, respectively, but at the same time it yields a more balanced model with respect to the predictivity of inhibitors and non-inhibitors (cf. Table 1).

**Construction of models by other machine learning methods**

We developed models with other methods to investigate if the quality of the models could be improved. For that purpose, classification models of active and inactive compounds were generated using the support vector machine (SVM), the kappa nearest neighbours (kNN), the random forest, and the decision tree methods.
As for the binary QSAR model described in the last section, care was taken with respect to what
descriptors that were included in the model building, to avoid overfitting and to reduce the noise.
We therefore generated models using all MOE and Volsurf variables, as well as using a variable
selection by the so-called BestFirst and a genetic algorithm (cf. Tables 2 and 3 and Figure 2). In
general, the quality of the models improved when using the BestFirst variable selection, whereas
variables selected by the genetic algorithm did not improve the model compared to using all
variables. For example, the total number of correctly predicted compounds in the test set is largest
using the BestFirst method (SVM\textsuperscript{D}: 73 %, SVM\textsuperscript{E}: 75 %, random forest: 76 %, kNN: 74 %,
decision tree: 71 %) (cf. Figure 2), and its ability to predict both actives and inactives is more
balanced compared to the models generated with all descriptors and descriptors selected by the
 genetic algorithm (cf. Table 3). Thus, in the remaining we will only comment on the models
generated by the BestFirst method for variable selection.

Overall, the different methods yield models of similar quality. Of the 7,058 compounds in the test
set, 4,984 – 5,375 (71 – 76 %) are correctly predicted. Thus, the different methods are quite similar
in performance, with SVM\textsuperscript{E} and the random forest methods being the best and the decision tree
method being a quite reasonable method. This is also reflected in the MCC values that are 0.51 –
0.52 for the SVM\textsuperscript{E} and the Random forest methods and 0.41 for the Decision tree method. It
should be noted, however, that the smallest discrepancy in prediction between training and tests is
observed for the model developed by the SVM\textsuperscript{D} method. This indicates that it is a well-balanced
and not overfitted model. As seen in Figure 2, the binary QSAR model is clearly worse than the
other methods, and again, this is also reflected in the MCC values of 0.31 and 0.36 (cf. Table 1).

A probability of belonging to a certain class, active or inactive, is given for each compound with
the random forest method. Compounds with a probability being smaller than 50 % are classified as
inactive compounds, and compounds with a probability larger than or equal to 50% are classified as active compounds. Figure 3 shows the percentage of true predictions for different ranges of the probabilities using the random forest model. Interestingly, there is a larger chance of finding either a true positive or a true negative for probabilities closest to 100% or 0%, respectively. For example, those compounds with a probability of 100% have the highest prediction rate of 95% which is significantly larger than if using lower values of the probability. Similarly, the compounds with a probability of 0% also have a larger prediction rate of 86% compared to higher values of the probability. This is an important result, because it gives a way of ranking the database and therefore also a way of selecting compounds from a database search.

External test set of drug compounds

In the following, we have focused on the best machine learning methods (SVM^D, SVM^E, random forest, kNN, and decision tree) to classify a small external set of 89 drugs (39 active and 50 inactive). Of these compounds, 24 violated Lipinski’s Rule-of-five, primarily due to lipophilicity. Figure 4 shows that this small external set of compounds span the same chemical space as the large PubChem dataset thus, it should be possible to predict their 1A2 activity with the developed models. The best method was again the SVM^E method, which classifies 60 of the 89 compounds correctly (cf. Table 4). Interestingly, the performance seems to be slightly poorer for this set of compounds as compared to the PubChem data set, which might be attributed to the smaller size of the set of compounds. Another explanation may arise from the fact that a different experimental assay was used to determine the inhibitory constants of the external test set when compared to the PubChem data set. This may lead to different experimental classifications for which the models were not trained.
Structural characteristics of inhibitors, non-inhibitors and misclassified compounds

It is important to understand why compounds are classified as they are. Generally, it turned out that differences in the hydrophobic and hydrophilic nature of the CYP1A2 compounds are important for discriminating inhibitors from non-inhibitors. In Volsurf, the hydrophobic and hydrophilic properties of a compound may be visualized with the so-called DRY (hydrophobic properties) and N1 (hydrogen donor properties) probes, as shown in Figure 5. Compound A, an inhibitor of CYP1A2, has a large hydrophobic region and a small polar region, which is common for most of the inhibitors. Non-inhibitors, as compound B, are characterized by slightly smaller hydrophobic regions and much larger polar regions. These characteristics fit well with the general idea of CYP1A2 ligands being hydrophobic with one hydrogen bond acceptor (Smith et al., 1997; Sansen et al., 2007). A compound with characteristics as compound C is likely to be predicted as a non-inhibitor, because it has a large polar region and therefore fails in the prediction. Similarly, compound D, which is a non-inhibitor, is mainly hydrophobic, and this is probably why this prediction also fails.

Lipinski-decision tree model

Although the models derived by the machine learning methods perform satisfactory, it would be useful to develop a simple "back-of-the-envelope calculation" like Lipinski’s Rule-of-Five. Such a method would be extremely fast and therefore potentially useful for high-throughput screening and at the same time allow an easy interpretation. Thus, a decision tree model was subsequently developed based on the four Lipinski parameters: molecular weight, number of hydrogen bond acceptors and donors, and the lipophilicity. The so-called Lipinski-decision tree model performed surprisingly well (cf. Table 5) 71 % and 67 % of the compounds in the training and test sets, respectively, were correctly predicted, even though the balance between TP and TN is relatively
far off. As illustrated in Figure 6 compounds classified as inhibitors are characterized as having either 1) more than 2 hydrogen bond donors, logP greater than or equal to 3.7 and less than 6 hydrogen bond donors, or 2) equal or less than 3 hydrogen bond donors and a molecular weight between 198 and 516. Compounds that fall outside these ranges are considered to be non-inhibitors. This decision model is in good agreement with the conclusion derived from inspection of the MIFs (cf. Figure 4) discussed in the previous section. Finally, we tested the Lipinski-decision tree model on a set of known 1A2 substrates and inhibitors (cf. Table 6) (Flockhart, 2007). Out of these, 34 compounds corresponding to 79% are correctly predicted as 1A2 ligands.
Conclusion

In this paper, we have reported the application of different machine learning classification methods like support vector machine (linear and non-linear), random forest, kNN, decision tree, and binary QSAR for the classification of cytochrome P450 1A2 ligands.

Models that are based on the support vector machine method are superior to binary QSAR models. Using the SVM, random forest, and kNN methods and the BestFirst variable selection method resulted in models with 73-76 % of the test set correctly predicted. Finally, a decision tree model based on Lipinski’s Rule-of-five descriptors was developed, classifying 67 % of the compounds in the test set correctly. This model is easy to interpret and offers structural insight into the classification of new CYP1A2 inhibitors.

Inspection of the MIFs illustrates the importance of hydrophobicity, hydrogen bond donors and acceptors. These features are in agreement with the Lipinski-decision tree model and agree with previously reported pharmacophore models and X-ray structure of the cytochrome P450 1A2 isoenzyme.

The models developed in this work are fast and precise enough to be applicable for virtual screening of large databases for identification of 1A2 inhibitors or non-inhibitors. Moreover, they can be used as filters to quickly assess the likelihood that a newly designed compound shows an interaction with CYP1A2. As such, the models can play an important role to prevent the risk of e.g. drug-drug interactions through metabolism at an early stage of the drug development process.
Acknowledgements

The authors are thankful to Professor Gabriele Cruciani, University of Perugia, Italy, for providing the Volsurf software.
References


The work has been supported by the Drug Research Academy, (P.V.), the NABIIT (Nanoscience Biotechnology and IT) program (O.T.), the Danish Medical Research Council (L.O.) and the Carlsberg Foundation (L.O.).
Legends for figures.

**Figure 1.**
Score plot from principal component analysis (two first components shown) of the PubChem data set. The compounds are coloured as follows: Training set and inactive as black diamonds, training set and active as red asterisk, test set and inactive as green triangles, and test set and active as blue circles.

**Figure 2.**
Correctly classified compounds from the test set in all methods. All variables: Predicted by all variables, BestFirst: Predicted by BestFirst variables, Genetic Search: Predicted by GeneticSearch variables, Max PC: Binary model using maximum number of principle components, Three PCs: Binary model using first three principle components. SVMD: Support vector machine, linear model, SVME: Support vector machine, non-linear model, RF: Random forest method, kNN: kappa nearest neighbors method, DT: Decision tree method, BQ: Binary QSAR

**Figure 3**
True predictions for probabilities of belonging to a certain class (data shown for random forest predictions of compounds in the test set). In the Weka software, compounds with a probability smaller or larger than 50% are classified as inactives or actives, respectively. Red/Blue: True predictions for inactive/active compounds.

**Figure 4**
The 89 drug molecules in the external test set projected on the score plot from principal component analysis of the PubChem data set.
**Figure 5**
GRID 3D molecular interaction fields (MIFs) of four representative compounds. The cyan contours (left column) correspond to MIFs calculated with an amide probe and contoured at -2 kcal/mol. The green contours (right column) correspond to MIFs calculated with a dry probe and contoured at -1 kcal/mol. From top to bottom the MIFs illustrate A) active compound classified as active, B) inactive compound classified as inactive, C) active compound classified as inactive, and D) inactive compound classified as active.

**Figure 6**
Decision tree model from Lipinski Rule-of-five descriptors (see text for details).
Table 1.

Binary QSAR models based on MOE and Volsurf descriptors.

<table>
<thead>
<tr>
<th>Set</th>
<th>Predicted compounds in %</th>
<th>Correctly predicted in %</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>Training</td>
<td>73</td>
<td>27</td>
<td>67</td>
</tr>
<tr>
<td>Training</td>
<td>96</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>Test</td>
<td>74</td>
<td>26</td>
<td>57</td>
</tr>
<tr>
<td>Test</td>
<td>55</td>
<td>45</td>
<td>80</td>
</tr>
</tbody>
</table>

*a* TP: Number of true positives, *FP*: Number of false positives, *TN*: Number of true negatives, *FN*: Number of false negatives. *b* Three principal components with a smoothing factor of 0.08. *c* Maximum principal components with a smoothing factor of 0.08 (default value in MOE). *d* MCC: Matthews correlation coefficient.
### Table 2.

Summary of results using all descriptors in training and test sets.

<table>
<thead>
<tr>
<th>Set</th>
<th>Method</th>
<th>Predicted compounds in %</th>
<th>Correctly predicted in %</th>
<th>MCC(^b)</th>
<th>5 fold CV(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
<td>FN</td>
</tr>
<tr>
<td>Training set</td>
<td>SVM(^d)</td>
<td>80</td>
<td>20</td>
<td>85</td>
<td>14</td>
</tr>
<tr>
<td>Training set</td>
<td>SVM(^e)</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Training set</td>
<td>RF(^f)</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Training set</td>
<td>kNN(^g)</td>
<td>84</td>
<td>16</td>
<td>72</td>
<td>28</td>
</tr>
<tr>
<td>Training set</td>
<td>C4.5/J48(^h)</td>
<td>96</td>
<td>4</td>
<td>97</td>
<td>3</td>
</tr>
<tr>
<td>Test set</td>
<td>SVM(^d)</td>
<td>80</td>
<td>20</td>
<td>61</td>
<td>39</td>
</tr>
<tr>
<td>Test set</td>
<td>SVM(^e)</td>
<td>77</td>
<td>23</td>
<td>67</td>
<td>33</td>
</tr>
<tr>
<td>Test set</td>
<td>RF(^f)</td>
<td>77</td>
<td>23</td>
<td>66</td>
<td>34</td>
</tr>
<tr>
<td>Test set</td>
<td>kNN(^g)</td>
<td>79</td>
<td>21</td>
<td>65</td>
<td>35</td>
</tr>
<tr>
<td>Test set</td>
<td>C4.5/J48(^h)</td>
<td>71</td>
<td>29</td>
<td>72</td>
<td>28</td>
</tr>
</tbody>
</table>

\(^a\) TP: Number of true positives, FP: Number of false positives, TN: Number of true negatives, FN: Number of false negatives. \(^b\) MCC: Matthews correlation coefficient. \(^c\) CV: Cross validation. \(^d\) SVM\(^D\): Support vector machine, linear model. \(^e\) SVM\(^E\): Support vector machine, non-linear model. \(^f\) RF: Random forest method. \(^g\) kNN: kappa nearest neighbors method. \(^h\) C4.5/J48: Decision tree method.
Table 3.

Summary of results using BestFirst descriptors in training and test sets.

<table>
<thead>
<tr>
<th>Set</th>
<th>Methods</th>
<th>Predicted compounds in %&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Correctly predicted in %</th>
<th>MCC&lt;sup&gt;b&lt;/sup&gt;</th>
<th>5 fold CV&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
<td>FN</td>
</tr>
<tr>
<td>Training</td>
<td>SVM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>74</td>
<td>26</td>
<td>79</td>
<td>21</td>
</tr>
<tr>
<td>Training</td>
<td>SVM&lt;sup&gt;e&lt;/sup&gt;</td>
<td>79</td>
<td>21</td>
<td>84</td>
<td>16</td>
</tr>
<tr>
<td>Training</td>
<td>RF&lt;sup&gt;f&lt;/sup&gt;</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>Training</td>
<td>kNN&lt;sup&gt;g&lt;/sup&gt;</td>
<td>86</td>
<td>14</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>Training</td>
<td>C4.5/J48&lt;sup&gt;h&lt;/sup&gt;</td>
<td>98</td>
<td>02</td>
<td>97</td>
<td>02</td>
</tr>
<tr>
<td>Test set</td>
<td>SVM&lt;sup&gt;d&lt;/sup&gt;</td>
<td>73</td>
<td>27</td>
<td>73</td>
<td>27</td>
</tr>
<tr>
<td>Test set</td>
<td>SVM&lt;sup&gt;e&lt;/sup&gt;</td>
<td>73</td>
<td>27</td>
<td>78</td>
<td>22</td>
</tr>
<tr>
<td>Test set</td>
<td>RF&lt;sup&gt;f&lt;/sup&gt;</td>
<td>78</td>
<td>22</td>
<td>74</td>
<td>26</td>
</tr>
<tr>
<td>Test set</td>
<td>kNN&lt;sup&gt;g&lt;/sup&gt;</td>
<td>79</td>
<td>21</td>
<td>68</td>
<td>32</td>
</tr>
<tr>
<td>Test set</td>
<td>C4.5/J48&lt;sup&gt;h&lt;/sup&gt;=</td>
<td>71</td>
<td>29</td>
<td>70</td>
<td>30</td>
</tr>
</tbody>
</table>

<sup>a</sup> TP: Number of true positives, FP: Number of false positives, TN: Number of true negatives, FN: Number of false negatives.  
<sup>b</sup> MCC: Matthews correlation coefficient.  
<sup>c</sup> CV: Cross validation.  
<sup>d</sup> SVM<sup>D</sup>: Support vector machine, linear model.  
<sup>e</sup> SVM<sup>E</sup>: Support vector machine, non-linear model.  
<sup>f</sup> RF: Random forest method.  
<sup>g</sup> kNN: kappa nearest neighbors method.  
<sup>h</sup> C4.5/J48: Decision tree method.
Table 4.
Prediction of external test set of 89 drug compounds using the machine learning methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Predicted compounds in %(^a)</th>
<th>Correctly predicted in %</th>
<th>MCC(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>SVM(^c)</td>
<td>56</td>
<td>44</td>
<td>70</td>
</tr>
<tr>
<td>SVM(^d)</td>
<td>62</td>
<td>38</td>
<td>72</td>
</tr>
<tr>
<td>RF(^e)</td>
<td>56</td>
<td>44</td>
<td>60</td>
</tr>
<tr>
<td>kNN(^f)</td>
<td>77</td>
<td>33</td>
<td>44</td>
</tr>
<tr>
<td>C4.5/J48(^g)</td>
<td>54</td>
<td>46</td>
<td>52</td>
</tr>
</tbody>
</table>

\(^a\) TP: Number of true positives, FP: Number of false positives, TN: Number of true negatives, FN: Number of false negatives. \(^b\) MCC: Matthews correlation coefficient. \(^c\) SVM\(^D\): support vector machine, linear model. \(^d\) SVM\(^E\): Support vector machine, non-linear model. \(^e\) RF: Random forest method. \(^f\) kNN: kappa nearest neighbors method. \(^g\) C4.5/J48: Decision tree method.
Table 5.
Result of Lipinski-decision tree model.

<table>
<thead>
<tr>
<th>Set</th>
<th>Predicted compounds in %&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Correctly predicted in %</th>
<th>MCC&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TP</td>
<td>FP</td>
<td>TN</td>
</tr>
<tr>
<td>Training set</td>
<td>91</td>
<td>09</td>
<td>53</td>
</tr>
<tr>
<td>Test set</td>
<td>87</td>
<td>13</td>
<td>43</td>
</tr>
</tbody>
</table>

<sup>a</sup> TP: Number of true positives, FP: Number of false positives, TN: Number of true negative, FN: Number of false negatives.<br>
<sup>b</sup> MCC: Matthews correlation coefficient.
### Table 6

Lipinski-decision tree model predictions of marketed CYP1A2 substrates/inhibitors (Flockhart, 2007).

<table>
<thead>
<tr>
<th>Prediction</th>
<th>1A2 substrates/inhibitors</th>
<th>Predictions in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>Amitriptyline, Ciprofloxacin, Clomipramine, Clozapine, Cyclobenzaprine, Estradiol, Fluvoxamine, Furafylline, Haloperidol, Imipramine, Levofloxacin, Methaxsalen, Mibefradil, α-Naphthoflavone, Naproxin, Narfloxacin, Olanzapine, Ondansetron, Propranolol, Riluzole, Ropivacaine, Tacrine, Tizanidine, Verapamil, Warfarin, Zileuton, Zolmitriptan</td>
<td>79</td>
</tr>
<tr>
<td>Wrong</td>
<td>Acetaminophen, Amiodarone, Caffeine, Cimetidine, Mexiletine, Phenacetin, Theophylline</td>
<td>21</td>
</tr>
</tbody>
</table>
Figure 1
Figure 2
Figure 3
Figure 4