Journal of Computer-Aided Molecular Design (2005) 19: 189–201
DOI 10.1007/s10822-005-3785-3

A support vector machine approach to classify human cytochrome P450 3A4 inhibitors

Jan M. Kriegl[a,*], Thomas Arnhold[b], Bernd Beck[a] & Thomas Fox[a]
[a]*Computational Chemistry, Department of Lead Discovery, Boehringer Ingelheim Pharma GmbH & Co. KG, D-88397 Biberach, Germany;* [b]*DDS-DMPK, Department of Drug Discovery Support, Boehringer Ingelheim Pharma GmbH & Co. KG, D-88397 Biberach, Germany*

Received 3 September 2004; accepted in revised form 14 March 2005

Key words: ADME, cytochrome P450, *in silico* filter, molecular descriptor, QSAR, support vector machine

Summary

The cytochrome P450 (CYP) enzyme superfamily plays a major role in the metabolism of commercially available drugs. Inhibition of these enzymes by a drug may result in a plasma level increase of another drug, thus leading to unwanted drug–drug interactions when two or more drugs are coadministered. Therefore, fast and reliable *in silico* methods predicting CYP inhibition from calculated molecular properties are an important tool which can be applied to assess both already synthesized as well as virtual compounds. We have studied the performance of support vector machines (SVMs) to classify compounds according to their potency to inhibit CYP3A4. The data set for model generation consists of more than 1300 structural diverse drug-like research molecules which were divided into training and test sets. The predictive power of SVMs crucially depends on a careful selection of parameters specifying the kernel function and the penalty for misclassifications. In this study we have investigated a procedure to identify a valid set of SVM parameters which is based on a sampling of the parameter space on a regular grid. From this set of parameters, either single SVMs or SVM committees were trained to distinguish between strong and weak inhibitors or to achieve a more realistic three-class assignment, with one class representing medium inhibitors. This workflow was studied for several kernel functions and descriptor sets. All SVM models performed significantly better than PLS-DA models which were generated from the corresponding descriptor sets. As a very promising result, simple two-dimensional (2D) descriptors yield a three-class model which correctly classifies more than 70% of the test set. Our work illustrates that SVMs used in combination with simple 2D descriptors provide a very effective and reliable tool which allows a fast assessment of CYP3A4 inhibition potency in an early *in silico* filtering process.

Abbreviations: ADME – absorption, distribution, metabolism and excretion; CYP – cytochrome P450; PLS – partial least squares; DA – discriminant analysis; SVM(s) – support vector machine(s); 3D – three-dimensional; 2D – two-dimensional; QM – quantum-mechanical; RBF – radial basis function.

Introduction

The cytochromes P450 (CYP) are a superfamily of heme-containing enzymes which play a major role in the metabolism of xenobiotics and endobiotics [1–3]. From this superfamily, members of the CYP1, CYP2, and CYP3 families primarily participate in the oxidative metabolism of drugs. The isoenzyme CYP3A4 is of special importance because it is the most abundant hepatic cytochrome P450 [3]. Furthermore, it is estimated that the CYP3A subfamily is involved in the

*To whom correspondence should be addressed. Fax: 49 7351 83 92237; E-mail: jan.kriegl@bc.boehringer-ingelheim.com

metabolism of more than 50% of all marketed drugs [4]. As it might lead to drug–drug interactions, inhibition of CYP3A4 should be investigated as early as possible in drug research. Although high-throughput *in vitro* assays are established which assess the inhibition of the metabolism of a particular P450 lead substrate in the presence of other compounds [5, 6], computational tools are highly desirable to evaluate compounds prior to synthesis. This is even more true in an early stage of a research project when several thousands or millions of compounds have to be characterized for lead identification [7]. Moreover, these tools can be utilized for instance to guide lead optimization or to prioritize further *in vitro* tests.

Recent years have seen enormous progress in understanding the molecular basis of CYP inhibition. The determination of P450 structures and site-directed mutagenesis experiments have provided insights into structure-function relationships of mammalian CYPs [8–14]. However, the *in silico* prediction of interactions between CYP3A4 and small organic compounds remains difficult. In the published crystal structures of CYP3A4 complexed with different substrates, different sizes and conformations of the binding site were found [12, 14], which complicates structure-based *in silico* approaches such as docking. As another option, three-dimensional (3D) QSAR and pharmacophore models of CYP3A4 substrates and inhibitors have been reported [15–17]. These models assume similar binding modes of all training set compounds, however. The large and flexible binding pocket of CYP3A4 should allow various ligand poses and conformations [18, 19]. Moreover, multiple substrate binding sites have been identified which may account for the structural diversity of known CYP3A4 inhibitors [20].

In this study we present an approach to classify CYP3A4 inhibitors which is based on a characterization of the ligands by various molecular descriptors. Ligand-based models for CYP3A4 inhibition using partial-least squares (PLS) [21], artificial neural networks [22], recursive partitioning [23], and, parallel to our study, support vector machines (SVMs) [24] have been reported.

SVMs are a very promising new methodology within the area of statistical learning theory and have been added to the data analysis toolbox in contemporary drug discovery [25]. Since their introduction, SVMs attracted attention in various areas, including bioinformatics [26] as well as molecular informatics and pharmaceutical research [27–32]. Here we describe a workflow which allows a fast and reliable assessment of CYP3A4 inhibition using SVMs. Since the performance of SVMs crucially depends on a set of parameters which have to be defined prior to the learning process, a strategy for parameter determination and model validation is one key component of the workflow. In addition, the capability of various molecular descriptors to classify compounds with respect to their CYP3A4 inhibition potency is investigated. These descriptors range from simple 2D descriptors to descriptors obtained after time consuming quantum-mechanical (QM) calculations. Models are derived which discriminate between strong and weak inhibitors or, in a practical more relevant application, which achieve a separation of strong, medium, and weak inhibitors. A comparison of the SVM based models with classifiers generated with standard PLS discriminant analysis (PLS-DA) provides further evaluation of the performance of the SVM approach.

Support vector machines

SVMs are based on the structural risk minimization principle from statistical learning theory and were originally designed for binary classification problems [25, 33]. In the case of linearly separable data, the SVM method finds the hyperplane $\mathbf{w}\cdot\mathbf{x} + b$ which yields the best discrimination between data points $\mathbf{x}$ of two distinct classes. Here, $\mathbf{w}$ denotes the normal vector of the hyperplane, and $b||\mathbf{w}||^{-1}$ is its perpendicular distance from the origin (see Figure 1). The corresponding decision function is given by $f(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x} + \text{b})$. The position and orientation of the hyperplane is adjusted such that the margin between both classes, $d = 2||\mathbf{w}||^{-1}$, is maximized. The decision function $f(\mathbf{x})$ is obtained after solving the constrained quadratic optimization problem and is given by

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{l} \tilde{\alpha}_i y_i (\mathbf{x}\cdot\mathbf{x}_i^{\text{sv}}) + \tilde{b}\right).$$

Note that only a subset of l training vectors called support vectors, $\mathbf{x}_i^{\text{sv}}$, contribute.

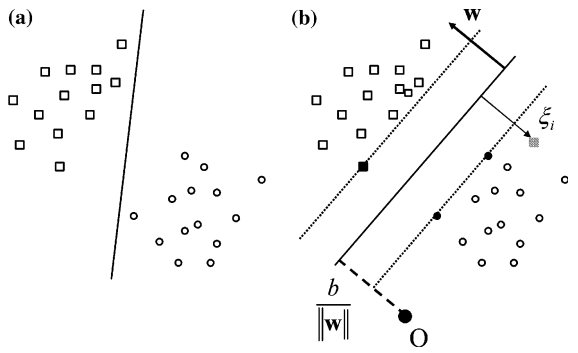

Figure 1. Principle of a support vector classifier. The solid lines in (a) and (b) separate the two identical sets of open circles and squares, representing two different classes. In contrast to (a), the solid line in (b) leaves the closest points (filled symbols) at the maximum distance. The margin, shown as dashed lines, is defined by the closest points, the so-called support vectors. The optimal separating hyperplane is defined by the normal vector, **w**, and the distance to the origin *O*, $b||\mathbf{w}||^{-1}$. In the case of non-separable sets, as indicated by the grey square among all open circles in (b), slack variables ξ_i are introduced which represent a relaxed margin.

If the classes overlap due to a high noise level or if there is a substantial degree of non-linearity between the attributes (e.g., descriptors) and the class membership, a perfect linear separation is not feasible. Two modifications of the linear approach address this problem. First, slack-variables ξ_i can be introduced which account for classification errors. The tradeoff between a maximum margin and the penalty for misclassification is regulated by a global parameter *C*. This corresponds to the introduction of a relaxed margin. Second, the data points can be transferred via a non-linear mapping into a higher-dimensional space where a linear separation is possible. Technically, the dot products $(\mathbf{x}_i \cdot \mathbf{x}_j)$ are replaced by their corresponding higher-dimensional expressions $k(\mathbf{x}_i, \mathbf{x}_j)$ (kernel function) rather than the original variables $\mathbf{x}_i$ and $\mathbf{x}_j$. We have used the Gaussian or radial basis function (RBF) kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2\right), \gamma \in R,$$

and the polynomial kernel,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \left((\mathbf{x}_i \cdot \mathbf{x}_j)\gamma + 1\right)^m, \gamma \in R, m \in N.$$

The extension from binary classification tasks to multiclass problems can be achieved with two different approaches: one is to construct and combine several binary classifiers, while in the other one all data are directly considered in one optimization problem. In this work, $k(k-1)/2$ binary classifiers are deduced from a training set with samples from k classes, and an unknown sample **x** is classified according to the class with the highest cumulative vote [34].

Materials and methods

CYP3A4 inhibition data

The data set consists of 1363 drug-like research compounds. These compounds were investigated in an assay determining the concentration causing a 50% inhibition of the turnover of a specific CYP3A4 substrate (IC_{50}). The data were determined with a radiometric assay using ^{14}C–N-methyl-erythromycin as substrate and recombinant CYP3A4 as the enzyme source [35]. The compounds were usually tested at physiologically relevant concentrations. Whether the mechanism of inhibition was competitive, non-competitive or uncompetitive was not investigated at that early stage.

After removal of outliers (see below), the final data set of 1345 molecules was divided into groups consisting of strong ($IC_{50} \leq 2$ μM; 243 molecules), medium ($2\ \mu M < IC_{50} \leq 20\ \mu M$; 561 molecules), and weak inhibitors ($IC_{50} > 20$ μM; 541 molecules).

Molecular descriptors

Four different descriptor sets have been applied in this study. Chemical properties based on the 2D structure of the molecule such as size, shape, lipophilicity, molar refractivity, atom and ring counts, and surface areas were calculated with an in-house program [36]. The final set consists of 32 2D descriptors (“in-house 2D descriptors”).

The second descriptor set is a subset of the QSAR descriptors available in MOE [37]. We selected 114 2D descriptors, including physical properties, surface areas, atom and bond counts, pharmacophore feature descriptors, and topological descriptors such as Kier–Hall connectivity indices or distance matrix descriptors (“MOE 2D

descriptors''). Subdivided surface area descriptors were not taken into account.

To assess the interaction energy of the molecule with its surrounding, we have used the VolSurf package [38], using four standard chemical probes (OH2, DRY, O, and N1) [39]. The 3D structure of the molecules was generated by first passing the 2D structure through Corina [40] and subsequently minimizing the conformational energy in the MMFF94s force field within MOE. From the 3D interaction maps, 88 descriptors were calculated (''VolSurf descriptors'').

The fourth descriptor set was generated with an in-house program based on quantum mechanical calculations using the AM1 Hamiltonian [41] within the semi-empirical program package VAMP 8.1 [42]. All these descriptors are, therefore, based on the 3D molecular structure. The resulting 68 descriptors cover electrostatic properties such as atomic charges derived from the electrostatic potential, the dipole moment, surface properties, the hydrogen bond behaviour, E-states, and other properties such as the molecular polarizability, the shape or the molecular volume (''in-house QM descriptors''). The calculation of these descriptors is most expensive compared to the other descriptor sets used in this study.

Of course, the performance of our models might be influenced by the preparation of the molecules before descriptor calculation and model development. First, we treated all ionizable centres in their neutral form. This does not necessarily correspond to the actual protonation state under physiological pH or within the active site of the enzyme. To our knowledge, there is no reliable method to calculate the protonation state of a ligand within the protein matrix which is also applicable to large data sets. Another issue in the preprocessing procedure is the generation of appropriate 3D conformers. As with the protonation state, the 3D conformation of the molecule within the active site of CYP3A4 is unknown and might be different from the conformation in an isotropic environment. In order to establish a workflow which can be applied automatically also to large (virtual) libraries, we put more emphasis in treating all compounds consistently rather than investigating the biological active conformation or the biological relevant ionization state.

Selection of training and test set

To incorporate the whole range of chemical and biological diversity into the model building process [43], a representative subset from the complete data set was selected for model training. We chose a ratio of ≈60% training data and 40% test data. The subset selection was achieved with a space-filling algorithm [21, 44], where the physico-chemical and biological properties of the compounds were represented by the scores of a PLS model between all descriptors and the log IC_{50} value.

Since one intention was to compare the performance of various descriptor sets, a unique training and a unique test set was used. However, the different numbers of descriptors within each set may lead to a biased subset selection if all descriptors were considered at a time in the PLS model. Therefore, we used a hierarchical approach to achieve a balanced selection [45, 46]. For each descriptor set, individual PLS models were generated. The score vectors of each individual base-level model comprise the X variables in the top-level model. Since the number of latent variables differs less than the number of descriptors, the influence of each descriptor set on the selection process is almost equal. Obvious outliers as determined from an examination of the score plot of the hierarchical PLS model were eliminated, because they would have been selected as training samples by the space-filling algorithm.

To obtain an additional and more objective estimation of the predictive power of our models, the entire data set was randomly split into multiple training and test sets. If not stated explicitly, the training and test set referred to in the following sections originate from the rational space-filling selection.

Multivariate data analysis

Basic and hierarchical PLS and PLS-DA was performed with the package SIMCA-P+ [47]. Data were mean-centered and scaled to unit variance before model generation. Descriptors with a variance below 0.0005 were eliminated. The number of significant latent components was determined by cross-validation [48].

Support vector machines

We have used the SVM software package LIBSVM 2.5 [49]. All SVM calculations were performed on an SGI cluster equipped with 24 SGI 600 MHz MIPS R14k processors (SGI, Mountain View, CA). The X variables of the training set were scaled to the interval [−1,1]. The same scaling factors were applied to the X variables of the test sets.

Cross-validation and determination of SVM parameters

We used the estimated predictive power of the classifier, q^2, as the objective function for parameter determination, where q^2 was determined by 10-fold cross-validation. In order to reduce the influence of the primary splitting step on the parameter selection procedure, q^2 was calculated for multiple random splits, thus providing a mean cross-validated model performance Q^2 and a corresponding standard deviation σ. Multiple random splits and noisy training data yield a rough Q^2-hypersurface with many local extrema. Hence, several cross-validation results have to be considered as equal within the statistical error bounds. Standard optimization algorithms such as gradient decent-like methods or simplex optimization will provide one local extremum of the cross-validation objective function representing one particular parameter set (C, γ) rather than an overview over all parameter combinations which yield feasible models. Therefore we applied a different strategy to determine the SVM parameters. The (C, γ) space was sampled on a regular grid [50]. At each grid point, the cross-validated performance of the correpsonding SVM classifier was computed. The standard deviation of the cross-validation results allows one to determine a minimum cross-validation performance which has to be exceeded. In this study, all parameter sets yielding cross-validation results within $\pm\sigma$ and $\pm 2.5\sigma$ of the maximum cross-validation result, $Q^2_{\max}$, were considered as valid models. They span the feasible SVM parameter space and were kept for further calculations as "set_1.0" and "set_2.5", respectively. The sampling was done on a logarithmic scale (logarithm with respect to base two) in order to cover a wide range of parameters. The range of $\log_2 C$ to be sampled was determined from the cross-validated performance of a linear classifier as suggested by Keerthi and Lin [50].

Although this heuristic grid search can be easily parallelized, it is computationally expensive. However, it yields a reliable estimation of the predictive power of a model which is not biased by one single subset splitting step. Furthermore, the influence of noise in the training data on the final model selection is reduced, since a set of models can be determined at this stage rather than the best cross-validated model.

In classification problems, the model performance can be evaluated as the fraction of properly classified items. To incorporate sensitivity and specificity, we calculated the generalized squared correlation coefficient C^2 [51]. In a multiclass prediction problem with k classes, all information about the model predictions is stored in a $k \times k$ contingency matrix $\mathbf{Z} = (z_{ij})$, where z_{ij} denotes all input elements predicted to be in class j while belonging to class i. The generalized squared correlation coefficient is given by

$$C^2 = \frac{1}{n(k-1)} \sum_{i,j} \frac{(z_{ij} - e_{ij})^2}{e_{ij}},$$

where n denotes the total number of elements to be classified, and $e_{ij} = (1/n) \sum_j z_{ij} \sum_i z_{ij}$ the expected number of data in cell i,j of the contingency matrix provided that there are no correlations between experimental assignments and predictions. For a binary classification problem, i.e. $k = 2$, the generalized squared correlation coefficient is equal to the squared Matthews correlation coefficient [51, 52]. The generalized squared correlation coefficient C^2 ranges from 0 to 1. A perfect prediction would lead to $C^2 = 1$.

All parameter pairs within the 2.5σ cut-off were used to train individual SVM classifiers. From the parameters pairs combined in set_1.0 and set_2.5, SVM committees rather than a single classifier were constructed [53]. The final prediction of the class membership was deduced

from the majority decision of all committee members.

Results and discussion

Selection of a representative training set

PLS models of all individual descriptor sets and hierarchical models were generated. The performance statistics of both the individual PLS models and the hierarchical models are summarized in Table 1. In contrast to the number of descriptors in each set, the number of significant latent components does not differ significantly; all descriptor sets contribute almost equally to the hierarchical model. After removal of 18 outliers from the data set, a refined hierarchical model was generated. Figure 2 shows the score plot of the first two principal components of the final hierarchical PLS-model. All five principal components of the refined hierarchical model were utilized for a representative selection of the training set by means of the space-filling algorithm. The training set of 807 molecules consists of 159 strong (20%), 325 medium (40%), and 323 weak inhibitors (40%). The complementary test set of 538 compounds contains 84 strong (16%), 234 medium (43%), and 220 weak inhibitors (41%). For binary SVM classifiers that were designed to distinguish between strong and weak inhibitors, all medium inhibitors were simply removed from the training and test set. The fraction of training data (482 compounds) is again close to 60%.

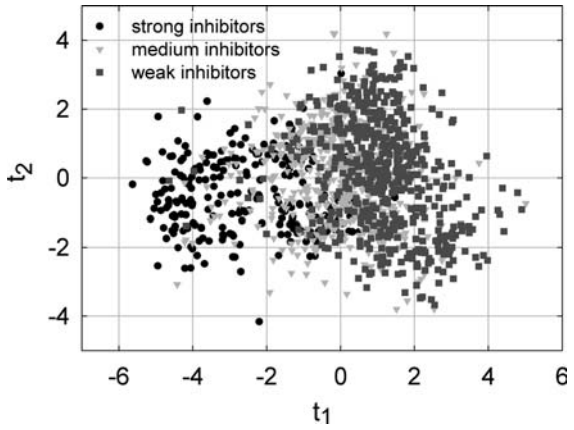

Figure 2. Score plot of the projection of the entire data set (1345 compounds) onto the first two latent components t_1 and t_2 of the hierarchical PLS model. Strong ($IC_{50} \leq 2$ μM), medium ($2\ \mu M < IC_{50} \leq 20\ \mu M$) and weak inhibitors ($IC_{50} > 20$ μM) are indicated by black, light grey, and dark grey symbols, respectively.

Determination of SVM parameters

The procedure to determine SVM parameters will be illustrated in the following with SVMs based on MOE 2D descriptors and, if not stated otherwise, with the RBF kernel. Similar results were obtained with the other descriptor sets.

Figure 3 shows the cross-validated squared correlation coefficient of a linear SVM classifier that separates between strong and weak inhibitors as a function of $\log_2 C$. From these results, we decided to restrict the grid search to $-5 \leq \log_2 C \leq 15$ to cover a reasonable range. The kernel parameter γ was limited to $-15 \leq \log_2 \gamma \leq 3$. The performance of the heuristic grid search was investigated for different numbers of random splits

Table 1. Performance statistics of PLS models.

Descriptor set	*N*	*K*	LC	R^2X	R^2Y	Q^2Y
MOE 2D	1363	91	6	0.78	0.47	0.43
In-house 2D	1363	32	6	0.81	0.42	0.38
VolSurf	1363	88	6	0.70	0.41	0.36
In-house QM	1363	64	7	0.67	0.41	0.36
All	1363	275	8	0.69	0.56	0.49
Hierarchical	1363	25	5	0.51	0.54	0.51
Refined hierarchical	**1345**	**25**	**5**	**0.51**	**0.54**	**0.51**

N = number of compounds; *K* = number of variables; LC = number of latent components; R^2X = fraction of explained *X*-variation; R^2Y = fraction of explained *Y*-variation; Q^2Y = fraction of the predicted *Y*-variation estimated with cross-validation. Bold: final model used for the selection of a representative training set.

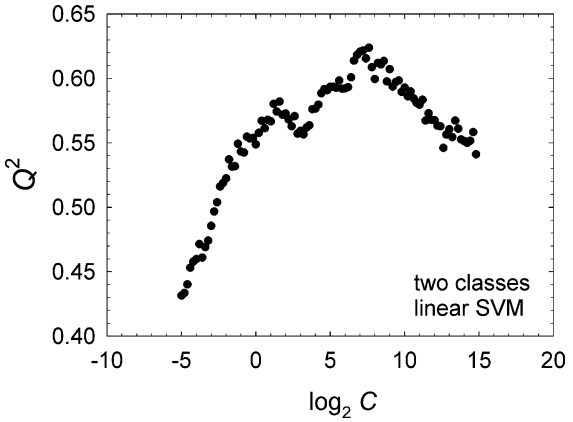

Figure 3. Cross-validated squared correlation coefficient Q^2 as a function of the error tradeoff C for linear SVMs which separate between strong and weak inhibitors. The calculations were performed with MOE 2D descriptors.

during cross-validation (25, 50, and 100) and two different resolutions Δ in $\log_2 C$ and $\log_2 \gamma$, namely $\Delta = 2.0$ and $\Delta = 0.2$ $\log_2$ units on both axes. The cross-validation results and the predictive power of the models did not depend significantly on the number of random splits. The higher grid resolution yielded slightly better cross-validation results. However, the models did not perform significantly better on the test set than models obtained in a low resolution grid search. Hence we conclude, that 25 different random splits and a grid resolution of $\Delta = 2.0$ provides a reliable way to determine valid SVM parameters.

As an example, Figure 4a shows the rough Q^2 surface for a binary classifier trained with strong and weak inhibitors. The grid resolution was set to $\Delta = 0.2$ $\log_2$ units in both $\log_2 \gamma$ and $\log_2 C$. This

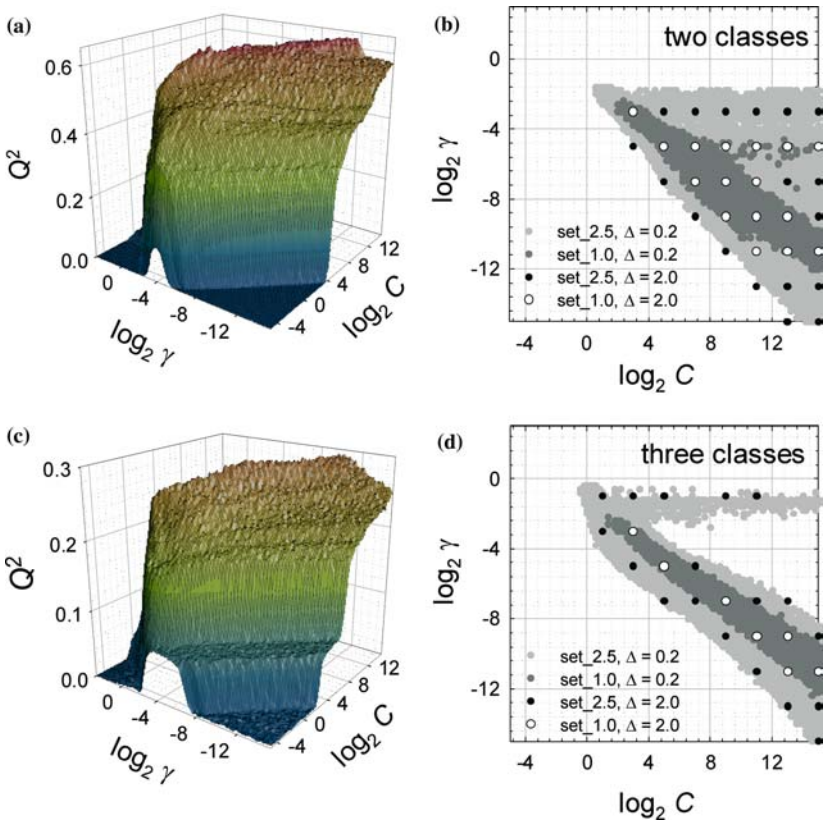

Figure 4. Determination of SVM parameters for the RBF kernel function. The surface plots (left) show the average cross-validated correlation coefficient Q^2 after 10-fold cross-validation using 25 different random splits of the training set. The resolution in both $\log_2 C$ and $\log_2 \gamma$ is $\Delta = 0.2$. On the right: parameter pairs of SVM classifiers with a cross-validation result Q^2 within σ (dark grey and open circles, set_1.0) and within 2.5σ (grey and black circles, set_2.5) of the maximum Q^2. The open and black circles represent parameters as determined from a grid search with $\Delta = 2.0$. All calculations were performed with MOE 2D descriptors. (a) and (b) two-class model; (c) and (d) three-class model. Note the different scales of the vertical axes in (a) and (c).

plot nicely shows the typical triangular shape of the parameter region with high cross-validation performance that has been deduced from the asymptotic behaviour of SVMs with a RBF kernel [50]. Sufficiently large values for the error tradeoff C together with small values for γ lead to severe overfitting, i.e. small regions around the training samples of the minority class are predicted to belong to that class whereas the rest of the data space is assigned to the majority class. In leave-group-out cross-validation, most members of the temporary validation set would then be classified according to the majority class, leading to a zero or close to zero squared correlation coefficient. For small error tradeoffs C and either very large or very small γ, the resulting SVM classifier tends to assign the entire data space to the majority class. Again, the corresponding cross-validated squared correlation coefficient is zero, as can be seen in the surface plot.

All parameter pairs yielding cross-validated performances within $Q^2_{max} \pm \sigma$ and $Q^2_{max} \pm 2.5\sigma$ are shown in Figure 4b as light and dark grey circles, respectively. For binary classification problems with little overlap, a considerable region of the (C, γ) space yields valid models. Parameter pairs as determined by the σ and 2.5σ cutoff after a low resolution grid search ($\Delta = 2.0$), plotted as filled and open circles in Figure 4b, still provide a good coverage of the region of interest. A similar distribution of feasible SVM parameters was found when a polynomial kernel was used instead of the RBF kernel (not shown).

The region of valid parameters becomes smaller if the classification problem is more complex and the classes overlap substantially. This is illustrated by the surface plot in Figure 4c, where the cross-validation results for a 3-class model that discriminates between strong, medium, and weak inhibitors, are shown. All valid parameter pairs above the σ and 2.5σ cut-off (same symbols as in Figure 4b) are depicted in Figure 4d.

The utility of the SVM committees derived from all valid parameter pairs is illustrated in Figure 5, where the relationship between the cross-validated squared correlation coefficient, Q^2, and the squared correlation coefficient determined from the test set predictions of each individual SVM classifier, C^2_{test}, is shown. These calculations were performed for the 3-class problem using a grid resolution of $\Delta = 0.2$. For models from set_1.0 (open diamonds), there is no correlation between Q^2 and C^2_{test} ($R^2 < 0.001$). Towards the lower bound of set_2.5 (open diamonds and filled circles), the quality of the test set predictions tends to become worse. For the whole set_2.5, the correlation coefficient between Q^2 and C^2_{test} is $R^2 = 0.32$. The comparison of the cross-validation results and the test set predictions demonstrates that the models in the committees have similar predictive powers. This is of relevance for practical applications where an independent test set which would allow to further validate the model and to perform a fine-tuning of the parameters is not available. Instead of arbitrarily defining the committee by the σ or 2.5σ cut off, one

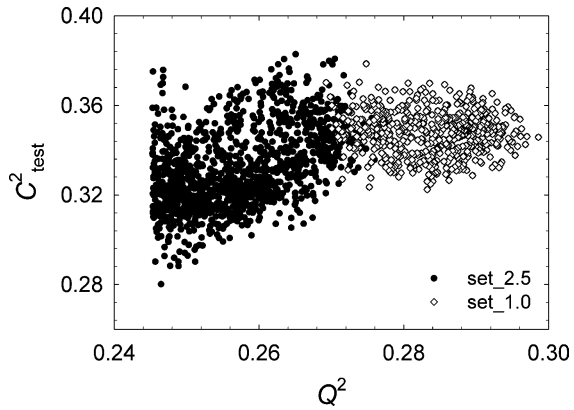

Figure 5. Comparison of the cross-validated squared correlation coefficient, Q^2, and the squared correlation coefficient of the test set predictions, C^2_{test}, for the committees set_1.0 (open diamonds) and set_2.5 (open diamonds and filled circles). The SVMs are based on MOE 2D descriptors, using the RBF kernel. The grids search was performed with $\Delta = 0.2$ in both $\log_2 C$ and $\log_2 \gamma$.

could also think to start with the best cross-validated model and increase the cut off until the squared correlation coefficient between Q^2 and C^2_{test} exceeds a user-defined threshold to get as many SVM classifiers as possible with equal predictive power.

Binary classifiers

Table 2 and Figure 6a summarize the best cross-validation results and the predictive power of the binary classification models generated from different descriptor sets and the representative training set. All models show an excellent separation of the test set compounds, with squared correlation coefficients often >0.7, corresponding to $>93\%$ correct class assignments. Similar results were obtained in other studies [21, 22, 24]. The different performance statistics of the individual models and the improved predictive power of a model constructed from all descriptors (see Table 2 and Figure 6a) indicate that the individual descriptor sets focus on slightly different aspects of the interaction between the ligand and the CYP3A4 active site. As an example, the 2×2 contingency table for the test set predictions from the committee set_2.5 using MOE 2D descriptors and the RBF kernel is visualized as bar chart in Figure 7a ($C^2 = 0.75$, total classification accuracy 94%). Table 2 also includes the performance statistics of SVM models generated with a polynomial and a linear kernel function. Models using non-linear kernel functions such as the RBF and the polynomial kernel perform better than a linear SVM, as can be expected from an introduction of additional degrees of freedom to adjust the separating hyperplane. Despite their convincing prediction performance, SVMs using a polynomial kernel will not be considered further in this study because of their longer training times compared to SVMs with a RBF kernel.

The comparison of all four models generated per descriptor set and kernel function again illustrates that the committees show a satisfying predictive power and often outperform the model trained with the best cross-validated parameters.

For comparison, test set predictions obtained from standard PLS-DA models are also included in Table 2 and Figure 6a. For all descriptor sets, SVMs yield significantly better results than the linear PLS-DA models. This is also illustrated in the bar chart in Figure 7b which shows the test set predictions from the PLS-DA model generated with the MOE 2D descriptors.

Multiclass models

The cross-validation results and the predictive power of all individual multiclass models are shown in Figure 6b (grid resolution $\Delta = 2.0$ on both axes) and Table 3. Depending on the descriptor set, the squared correlation coefficient C^2 of the test set predictions is between 0.3 and 0.44, corresponding to overall classification accuracies between 70 and 75%. These accuracies are also obtained from models that were trained using a randomly chosen training set. For example, with MOE 2D descriptors, the complementary test set is

Table 2. Performance statistics of binary SVM classifiers.

descriptor set	kernel	Q^2_{max}	top cv	top	set_1.0	set_2.5	PLS-DA
MOE 2D	RBF	0.63 ± 0.02	0.65	0.75	0.74	0.75	0.53
MOE 2D	RBF[a]	0.64 ± 0.02	0.65	0.77	0.68	0.74	
MOE 2D	polynomial[b]	0.63 ± 0.02	0.66	0.75	0.64	0.72	
MOE 2D	linear	0.62 ± 0.03	0.62	0.69	0.62	0.64	
in-house 2D	RBF	0.60 ± 0.03	0.57	0.71	0.63	0.64	0.56
VolSurf	RBF	0.52 ± 0.03	0.47	0.64	0.59	0.67	0.43
in-house QM	RBF	0.52 ± 0.03	0.66	0.73	0.73	0.73	0.36
all	RBF	0.67 ± 0.02	0.73	0.74	0.72	0.73	0.52

[a]resolution: $\Delta = 0.2$ in both C and γ.
[b]degree 5.
Q^2_{max} = best cross-validation result; top cv = model trained with the best cross-validated parameter pair ($\log_2 C$, $\log_2 \gamma$); top = model with the best test set prediction; set_1.0 and set_2.5 = SVM committees derived from parameter pairs which yield cross-validation results within the σ and 2.5σ cut off, respectively.

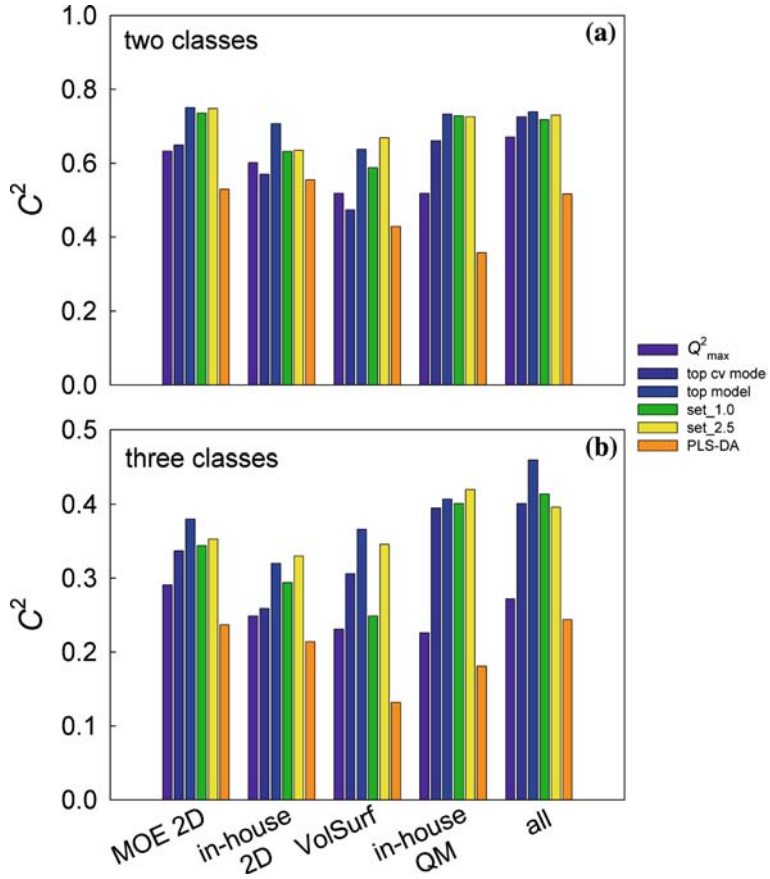

Figure 6. Comparison of different descriptor sets according to their cross-validated squared correlation coefficient and the results of test set predictions using the best cross-validated model (top cv model), the model with the best test set predictions (top model), and the committees set_1.0 and set_2.5, respectively. For comparison, also the test set predictions of the corresponding PLS-DA models are included. (a) strong and weak inhibitors. (b) strong, medium, and weak inhibitors. Note the different scales of the vertical axes.

correctly re-classified with 70% accuracy and $C^2 \approx 0.3$ (10 different random splits).

The complex multiclass decision is much more sensitive upon the kind of chemical information which is encoded by the individual descriptor set than the discrimination of strong and weak inhibitors. Here, the best test set predictions are obtained with the in-house QM descriptors. Models that were constructed from the in-house 2D and the VolSurf descriptors are less predictive. The MOE 2D descriptors yield classifiers which correctly classify 71% of the whole test set, with a squared correlation coefficient close to 0.4. This is a very promising result since the MOE 2D descriptors are fast to compute and hence allow also to make predictions for large compound collections within an acceptable period of time. A more detailed insight into the prediction performance is provided by an analysis of the contingency table. Figure 7c illustrates the test set predictions that were obtained with MOE 2D descriptors and the committee set_2.5. 63% of the weak inhibitors and 68% of the strong inhibitors are correctly recognized by this classifier. Moreover, almost none of the strong inhibitors are classified as weak inhibitors and *vice versa*. This was also observed for models derived from the other descriptor sets. As in the binary classification problem, the committees are able to correctly re-classify the compounds of the test set comparably or even better than the models trained with the best cross-validated parameters.

The predictive power of SVM classifiers is again illustrated by the comparison with 3-class PLS-DA models. Although the PLS-DA models still yield acceptable overall classification

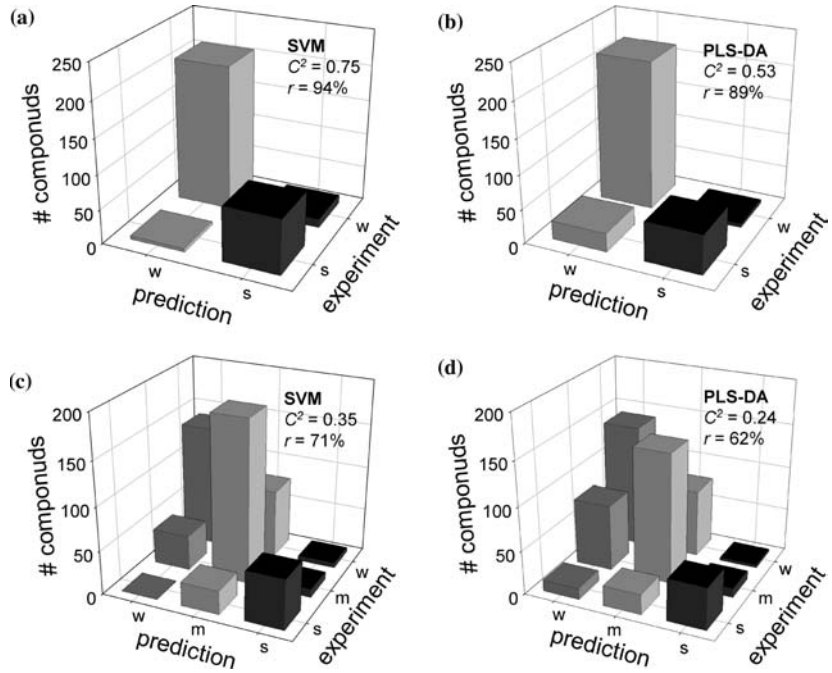

Figure 7. Comparison of experimental and predicted class memberships for the test set according to the committee set_2.5 (a, c) and the PLS-DA model (b, d). All models were derived from MOE 2D descriptors. (a) two-class model. (b) three-class model, respectively. s = strong, m = medium, w = weak inhibitors. C^2 = squared correlation coefficient, r = overall accuracy.

accuracies (e.g., 62% with the MOE 2D descriptors), they are clearly less predictive than the corresponding SVM models. Especially the strong inhibitors are more often wrongly classified, as illustrated in the bar chart representation of the test set predictions in Figure 7d.

The final test for every predictive model is its blind application to new, previously unseen data. This corresponds to the application scenario of QSAR models in a research project. As an example, we have applied both the binary and the multiclass SVM committee derived from the MOE 2D descriptors to seven drugs with CYP3A4 inhibition potency known from the literature [54] (see Figure 8). This small collection contains only molecules which can be unambiguously defined as either strong or weak inhibitor. The prediction results summarized in Table 4 show that all weak inhibitors are classified correctly, and only one strong inhibitor (nicardipine) is wrongly classified as weak inhibitor.

Table 3. Performance statistics of 3-class models.

descriptor set	kernel	Q^2_{max}	top cv	top	set_1.0	set_2.5	PLS-DA
MOE 2D	RBF	0.29 ± 0.01	0.34	0.38	0.34	0.35	0.25
MOE 2D	RBF[a]	0.30 ± 0.01	0.35	0.38	0.36	0.36	
MOE 2D	RBF[b]	0.28 ± 0.01	0.31	0.34	0.30	0.31	0.23
in-house 2D	RBF	0.25 ± 0.01	0.26	0.32	0.29	0.33	0.25
VolSurf	RBF	0.23 ± 0.01	0.31	0.37	0.25	0.35	0.16
in-house QM	RBF	0.23 ± 0.01	0.40	0.41	0.40	0.42	0.15
all	RBF	0.27 ± 0.01	0.40	0.46	0.41	0.40	0.27

[a]resolution: $\Delta = 0.2$ in both $\log_2 C$ and $\log_2 \gamma$.
[b]randomly splitted data set (average of 10 random splits).

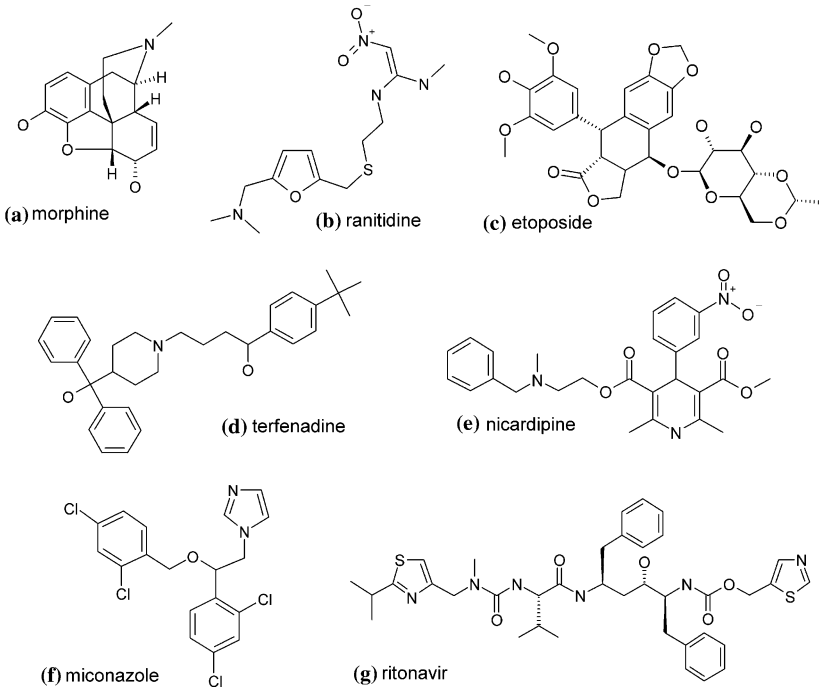

Figure 8. Seven known drugs that have been used in the blind test (see Table 4).

Conclusion

In this study we have investigated a support vector approach to classify compounds according to the inhibition of CYP3A4. We have shown that a heuristic search over the SVM parameter space in combination with extensive cross-validation is a valuable approach to determine key SVM parameters and to estimate the predictive power of the individual classifiers. Moreover, this procedure allows to generate SVM committees which yield reliable predictions. Our binary SVM models show a similar performance as reported in a recent study [24]. However, we extended the methodology also to discriminate between strong, medium, and weak CYP3A4 inhibitors. Very promising results were obtained with simple and fast 2D descriptors. The best models obtained from 2D descriptors were able to distinguish between strong, medium, and weak inhibitors of a test set with an accuracy of more than 70%. This is of great practical relevance since 2D descriptors are fast to compute. Together with the prediction speed of the SVM classifier, the CYP3A4 inhibition potency of even large compound libraries can be evaluated. As demonstrated in our approach to generate and apply SVM classifiers, our SVM models have to be considered as filtering tools rather than models which provide a direct and interpretable link between the chemical structure and the biological activity. However, their predictive power and robustness against overfitting makes them an essential tool in an early stage of a drug discovery project. In a recent study, Byvatov and Schneider introduced a SVM-based feature selection which allows one to identify the most relevant descriptors [55]. The possibility to extract structural information from an SVM classifier will further enhance the applicability and acceptability of this method.

Table 4. Classification of known drugs with the 2-class and 3-class SVM committees set_2.5 based on MOE 2D descriptors.

name	experiment	two-class	three-class
(a) morphine	w	w	w
(b) ranitidine	w	w	w
(c) etoposide	w	w	w
(d) terfenadine	s	s	m
(e) nicardipine	s	w	w
(f) miconazole	s	s	s
(g) ritonavir	s	s	s

[a]according to Ref. [54].
w, m, s = weak, medium, strong inhibitor.

Acknowledgement

We would like to thank Professor Chih-Jen Lin (National Taiwan University) for kindly allowing us to use LIBSVM and for helpful comments in the early stages of this study.

References

1. Lewis, D.F., Curr. Med. Chem., 10 (2003) 1955.
2. Danielson, P.B., Curr. Drug Metab., 3 (2002) 561.
3. Rendic, S. and Di Carlo, F.J., Drug Metab. Rev., 29 (1997) 413.
4. Wrighton, S.A., Schuetz, E.G., Thummel, K.E., Shen, D.D., Korzekwa, K.R. and Watkins, P.B., Drug Metab. Rev., 32 (2000) 339.
5. Miller, V.P., Stresser, D.M., Blanchard, A.P., Turner, S. and Crespi, C.L., Ann. NewYork Acad. Sci., 919 (2000) 26.
6. Jenkins, K.M., Angeles, R., Quintos, M.T., Xu, R., Kassel, D.B. and Rourick, R.A., J. Pharm. Biomed. Anal., 34 (2004) 989.
7. Böhm, H.-J. and Schneider, G., Virtual Screening for Bioactive Molecules. Wiley-VCH, New York, 2000.
8. Schoch, G.A., Yano, J.K., Wester, M.R., Griffin, K.J., Stout, C.D. and Johnson, E.F., J. Biol. Chem., 279 (2004) 9497.
9. Szklarz, G.D. and Halpert, J.R., Drug Metab Dispos., 26 (1998) 1179.
10. Wester, M.R., Johnson, E.F., Marques-Soares, C., Dansette, P.M., Mansuy, D. and Stout, C.D., Biochemistry, 42 (2003) 6370.
11. Williams, P.A., Cosme, J., Ward, A., Angove, H.C., Vinković, D.M. and Jhoti, H., Nature, 424 (2003) 464.
12. Williams, P.A., Cosme, J., Vinković, D.M., Ward, A., Angove, H.C., Day, P.J., Vonrhein, C., Tickle, I.J. and Jhoti, H., Science, 305 (2004) 683.
13. Wester, M.R., Yano, J.K., Schoch, G.A., Yang, C., Griffin, K.J., Stout, C.D. and Johnson, E.F., J. Biol. Chem., 279 (2004) 35630.
14. Yano, J.K., Wester, M.R., Schoch, G.A., Griffin, K.J., Stout, C.D. and Johnson, E.F., J. Biol. Chem., 279 (2004) 38091.
15. Ekins, S., Bravi, G., Wikel, J.H. and Wrighton, S.A., J. Pharmacol. Exp. Ther., 291 (1999) 424.
16. Ekins, S., Stresser, D.M. and Williams, J.A., Trends Pharmacol. Sci., 24 (2003) 161.
17. Ekins, S., Bravi, G., Binkley, S., Gillespie, J.S., Ring, B.J., Wikel, J.H. and Wrighton, S.A., J. Pharmacol. Exp. Ther., 290 (1999) 429.
18. Kumar, G.N. and Surapaneni, , Med. Res. Rev., 21 (2001) 397.
19. Szklarz, G.D. and Halpert, J.R., J. Comput.-Aided Mol. Des., 11 (1997) 265.
20. Schrag, M.L. and Wienkers, L.C., Arch. Biochem. Biophys., 391 (2001) 49.
21. Zuegge, J., Fechner, U., Roche, O., Parrott, N.J., Engkvist, O. and Schneider, G., Quant. Struct. Act. Relat., 21 (2002) 249.
22. Molnár, L. and Keseru, G.M., Bioorg. Med. Chem. Lett., 12 (2002) 419.
23. Ekins, S., Berbaum, J. and Harrison, R.K., Drug Metab. Dispos., 31 (2003) 1077.
24. Merkwirth, C., Mauser, H., Schulz-Gasch, T., Roche, O., Stahl, M. and Lengauer, T., J. Chem. Inf. Comput. Sci., 44 (2004) 1971.
25. Vapnik, V. The Nature of Statistical Learning Theory. Springer, New York, 1995.
26. Lee, Y. and Lee, C.K., Bioinformatics, 19 (2003) 1132.
27. Byvatov, E., Fechner, U., Sadowski, J. and Schneider, G., J. Chem. Inf. Comput. Sci., 43 (2003) 1882.
28. Warmuth, M.K., Liao, J., Rätsch, G., Mathieson, M., Putta, S. and Lemmen, C., J. Chem. Inf. Comput. Sci., 43 (2003) 667.
29. Trotter, M.W.B. and Holden, S.B., Quant. Struct. Act. Relat., 22 (2003) 533.
30. Zernov, V.V., Balakin, K.V., Ivaschenko, A.A., Savchuk, N.P. and Pletnev, I.V., J. Chem. Inf. Comput. Sci., 43 (2003) 2048.
31. Lind, P. and Maltseva, T., J. Chem. Inf. Comput. Sci., 43 (2003) 1855.
32. Sorich, M.J., Miners, J.O., McKinnon, R.A., Winkler, D.A., Burden, F.R. and Smith, P.A., J. Chem. Inf. Comput. Sci., 43 (2003) 2019.
33. Cortes, C. and Vapnik, V., Mach. Learn., 20 (1995) 273.
34. Hsu, C.-W. and Lin, C.-J., IEEE Transactions on Neural Networks, 13 (2002) 415.
35. Moody, G.C., Griffin, S.J., Mather, A.N., McGinnity, D.F. and Riley, R.J., Xenobiotica, 29 (1999) 53.
36. These descriptors are calculated by a Boehringer Ingelheim in-house software package (propty, developed by K.M. Hasselbach).
37. Molecular Operating Environment Release 2003.2, Chemical Computing Group, Montreal, Canada, 2003.
38. VolSurf 3.0.11, Molecular Discovery Ltd., London, UK, 2004.
39. Cruciani, G., Pastor, M. and Guba, W., Eur. J. Pharm. Sci., 11 (Suppl 2) (2000) S29.
40. CORINA 3.1, Molecular Networks GmbH, Erlangen, Germany, 2004.
41. Dewar, M.J.S., Zoebisch, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
42. VAMP 8.1, University of Erlangen, Erlangen, Germany (This version is provided as part of Materials Studio 2.2.1 by Accelrys, Inc.), 2003.
43. Golbraikh, A., Shen, M., Xiao, Z., Xiao, Y.D., Lee, K.H. and Tropsha, A., J. Comput.-Aided Mol. Des., 17 (2003) 241.
44. Kennard, R.W. and Stone, L.A., Technometrics, 11 (1969) 137.
45. Eriksson, L., Johansson, E., Lindgren, F., Sjøstrøm, M. and Wold, S., J. Comput.-Aided Mol. Des., 16 (2002) 711.
46. Eriksson, L., Arnhold, T., Beck, B., Fox, T., Johansson, E. and Kriegl, J.M., J. Chemometrics, 18 (2004) 188.
47. SIMCA-P+ 10, Umetrics AB, Umeå, Sweden, 2004.
48. Wold, S., Technometrics, 20 (1978) 397.
49. LIBSVM 2.5 National Taiwan University, 2003; http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.
50. Keerthi, S.S. and Lin, C.-J., Neural Comput., 15 (2003) 1667.
51. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. and Nielsen, H., Bioinformatics, 16 (2000) 412.
52. Matthews, B.W., Biochim. Biophys. Acta, 405 (1975) 442.
53. Bishop, C. M. and Bishop, C. M., Br. J. Clin. Pharmacol., 57 (2004) 473.
55. Byvatov, E. and Schneider, G., J. Chem. Inf. Comput. Sci., 44 (2004) 993.