Classification of Highly Unbalanced CYP450 Data of Drugs Using Cost Sensitive Machine Learning Techniques

T. Eitrich,[†] A. Kless,*,[‡] C. Druska,[†] W. Meyer,[†] and J. Grotendorst[†]

Central Institute for Applied Mathematics, Research Centre Jülich, Germany, and Grünenthal GmbH, Drug Discovery, Aachen, Germany

In this paper, we study the classifications of unbalanced data sets of drugs. As an example we chose a data set of 2D6 inhibitors of cytochrome P450. The human cytochrome P450 2D6 isoform plays a key role in the metabolism of many drugs in the preclinical drug discovery process. We have collected a data set from annotated public data and calculated physicochemical properties with chemoinformatics methods. On top of this data, we have built classifiers based on machine learning methods. Data sets with different class distributions lead to the effect that conventional machine learning methods are biased toward the larger class. To overcome this problem and to obtain sensitive but also accurate classifiers we combine machine learning and feature selection methods with techniques addressing the problem of unbalanced classification, such as oversampling and threshold moving. We have used our own implementation of a support vector machine algorithm as well as the maximum entropy method. Our feature selection is based on the unsupervised McCabe method. The classification results from our test set are compared structurally with compounds from the training set. We show that the applied algorithms enable the effective high throughput in silico classification of potential drug candidates.

1. INTRODUCTION

The complex process of identification of new potential drug candidates is not only related to find the most potent compound for a drug target. Through experience from the past, we know that the process of drug metabolism and pharmacokinetics should be considered equally in selecting new compounds. In this paper, we focus on a small part of these potential problems addressing the ADMET properties of a drug[1−3] which are related to the cytochrome P450 class of enzymes. Cytochrome P450s are drug metabolizing enzymes that among various other tissues have their primary site of drug metabolism in the liver. Herein the cytochrome P450 (CYP) superfamily plays a major role in degradation through oxidation of drugs. It has been reported that the five most important isoforms of the cytochrome P450 superfamily are 1A2, 2C19, 2C9, 2D6, and 3A4.[4,5] These are involved in the metabolism of about 90% of all drugs. According to Flockhart[6] inhibitors block the function of the enzyme so that its overall activity is reduced. Recently[7] we focused on the 1A2 isoform, whereas the 2D6 isoform is even more important since it is not active in some parts of the population resulting in higher plasma levels of 2D6 substrates. As could be learned from pharmacophor and CYP2D6 active sites studies, molecules with a basic nitrogen atom and an aromatic moiety are often recognized.[1−3,8−13] Additionally, if one compound inhibits CYP2D6, the subsequent decrease of metabolism of another drug can lead to unexpected drug−drug interactions.[14] This is due to accumulation of the latter compound as it is not being metabolized. Therefore, inhibi-

tion of CYP2D6 is an undesirable feature in a drug candidate. The ability to predict the inhibition of this enzyme starting from a new chemical entity (NCE) is important because obtaining high quality experimental data under in vivo conditions is time-consuming, has low throughput, is resource demanding, and therefore only available at a late stage of the drug discovery process. Therefore the in silico prediction of drug metabolism profiles of CYP450s has become one of the key technologies in early drug discovery support.[15−24]

The primary aim of this paper is to demonstrate how to build useful classification models out of unbalanced[25−28] data sets. We consider a data set to be unbalanced if either the sizes of the two classes differ significantly, or the costs for a false negative classification are very high whereas a false positive is acceptable, or if both conditions hold. Various classical approaches have been applied to understand the mechanism of CYP450 2D6 inhibition, like quantum mechanical calculations, pharmacophore modeling, and protein homology modeling.[11−14,20−22] There is an increasing need for robust and accurate classification algorithms that support the high throughput analysis of virtual libraries of molecules. Many working groups have developed prediction systems using the whole arsenal of computational methods to address the problem of CYP450 classification of NCEs.[1−3,18] It has been discovered that a combination of different classifications into a consensus model can improve the overall results of the classification. The basic algorithms behind were decision tree methods like recursive partitioning, ensemble methods[22,23,29−32] like adaptive boosting,[33] and support vector machines.[34] However, the limiting but most important part is the access to validated data. Usually the generalization behavior of a classifier improves with the amount of data points that are used in the model. But this does not necessarily mean that

* Corresponding author phone: +49 241 5692378; e-mail: Achim.Kless@grunenthal.com.
† Research Centre Jülich.
‡ Grünenthal GmbH.

a collection from various data sources representing different experiments will reveal the best classification. With the focus on data quality there is need for the use of unbalanced data sets because the number of known 2D6 inhibitors is limited. Conversely the reduction of noninhibitors to generate a balanced data set is accompanied by information loss. Similar to the number of instances is the number of molecular descriptors where irrelevant molecular descriptors reduce the prediction accuracies. In this paper, we show that our applied computational methods are able to predict the metabolic properties of NCEs for the virtual screening of large compound libraries. This enables the support for the design of compounds with desirable metabolic behavior.

Based on machine learning algorithms for classification, a framework was developed for cost sensitive learning by integrating several components that prevent the model from simply learning the simplest rule with a low rate of true positive classifications. We use the learning methods called support vector machine (SVM)[35] and maximum entropy model (ME).[36] To further improve the recognition abilities of these methods we analyze the influence of feature selection methods as well as techniques that are aimed at boosting sensitivity and accuracy. A challenging problem in the classification of unbalanced data sets is the aspect of overfitting. When models become too powerful on the training set, they may be useless for the classification of unseen data. We treat this problem by strictly dividing all data sets into training and test sets where the latter ones are not shown to the classifier during the learning process.

The remainder of this paper is organized as follows. In section 2, we describe the analyzed data sets. Our machine learning framework for cost sensitive classification will be introduced in section 3. In section 4, we present our experimental results. Finally, section 5 contains a summary and shows directions for future work.

2. DESCRIPTION OF THE DATA SETS

The application of our algorithms needs to be done in feature vectors where every chemical structure is represented by a few hundred physicochemical features or substructural features that are summarized in binary form.[37] The feature vectors are normalized representions of the chemical structure by a similar count of features that are independent from the size of the molecules. In the second step, we add the specific information of the 2D6 inhibition that is known. We repeat this step for every molecule in the data set so that we finally get a two-dimensional relation between the properties and the 2D6 inhibition. The generated data sets are cleaned by removing redundant information like zero columns. Based on 2D6 inhibition data from 263 structurally diverse drugs or druglike molecules which we extracted from publically available data,[4−6] we split the data set into a diverse set of 185 compounds for training and 78 compounds that we used as the test set.[38−41] A drug was selected as 2D6 inhibitor if there is published evidence that the compound is a 2D6 inhibitor, which was decided on the basis of either in vitro assay with recombinant 2D6 or under physiological conditions with hepatocytes or liver microsomes. It does not necessarily follow that the isoform is the principal metabolic pathway in vivo or that alterations in the rate of the metabolic reaction catalyzed by that isoform will have large effects on the pharmacokinetics of the drug. Such a data set compiled from diverse public sources may vary to some extent on conditions used in the assays, e.g., use of different labeled substrates like dextromethorphan. We pooled potent inhibitors with an $IC_{50} < 2$ μM together with weak inhibitors up to 50 μM since we are using binary classifiers. All compounds from our data set are well-known drugs, and therefore we assume that a relevant 2D6 inhibition would have been published. For these cases, we classified the corresponding compounds to be noninhibitors. The training set of 185 compounds contains 35 compounds that inhibit the CYP2D6 isoform and 150 that do not. We refer to the former as positive training points (class 1) and the latter as negative training points (class −1). The test set contained the same ratio of positive to negative molecules as the training set. For the oversampling method[42] we increased the amount of positives in the training set with a factor of 3 by repeating each feature vector three times. The split of the presented compounds into training set and test set was carried out on the basis of their Tanimoto coefficients representing chemical similarity[38] and diversity.[39] For that purpose, the pairwise Tanimoto coefficients matrix has been generated during the clustering procedure within the MOE program package using the bit packed MACCS structural keys. Whereas the direct interpretation of chemical structures is best done by visual inspection of the classification results, for building classifiers, we need to describe the chemical structures by calculation of physicochemical features. These molecular descriptors are mathematical representations of the chemical structures. For this purpose we were using the properties as implemented in the MOE[43] program package.

Overall, we have calculated 557 ensemble descriptors including easily interpretable descriptors like atom-count descriptors,[44] others which describe the connectivity of the molecular graph,[45] like constitutional descriptors,[46] the topological descriptors of Randic[47] or Kier and Hall,[48,49] autocorrelations, and aromaticity indices[50] as well as 3D-descriptors[51] that encode the 3D van der Waals surface of the molecule annotated with physical properties like charge, hydrogen bond, or acceptor potential. The 3D-based features were calculated from the 3D coordinates of the compounds generated by Corina.[52−54] The calculation of a 458 bit length screening vector (E_Screen) of the chemical structures was done with the CACTVS program.[55]

Overall, we have generated four data sets. The ensemble data set contains only the ensemble features and leads to one feature vector per drug. The binary data set contains the substructural features in binary form. In case of the ME algorithm, we had to modify the feature vectors because the algorithm is based on categorical features. In case of the ensemble vectors, we were using an unsupervised discretization procedure where every distribution of the normalized feature is split into 10 equal bins. In a second step, we proceeded as we did for the binary data set in such a way that the name of each feature was connected to the corresponding bin of the discretization forming a categorical feature. For the binary data set, each bin had to be named and combined with the corresponding binary value. This results in categorical feature vectors where every bin feature can be distinguished from the other. The test set always remained the same for the ensemble as well as the binary data set and was used consistently for all classification efforts in this paper.

3. MACHINE LEARNING TECHNIQUES

Machine learning (ML) is said to be the development of computer programs which allow to learn rules by analysis of data sets. Many other interpretations of machine learning do exist. Our work is concerned with a subarea of ML called supervised learning.[35] We concentrate on learning hypothesis functions for binary classification problems.[56] Given a training set

$$\{(x^i, y_i) \in \mathbb{R}^n \times \{-1, 1\}, i = 1,\ldots,l\} \qquad (1)$$

of $l \in \mathbb{N}_+$ input−output pairs, each with $n \in \mathbb{N}$ real valued attributes and a binary class label y_i, the task of supervised learning is to find a hypothesis function $h : \mathbb{R}^n \rightarrow \{-1; 1\}$ that can be used to classify unseen data. Several approaches for learning h have been proposed, such as neural networks, decision trees, and nearest neighbor methods.[57] Among the modern ML methods, the support vector machine approach turned out to be reliable and powerful.[34] Based on a standard SVM implementation, we modified the algorithm to solve classification problems on noisy and unbalanced data. We also studied the maximum entropy framework for classification. Usually applied to problems of natural language processing, we examined the benefit of the ME classification principle for data sets of chemical structures. The freely available *openNLP* maximum entropy software package was applied.[58] In this paper which is aimed at the classification of unbalanced data, we describe our work on SVMs and ME models. In addition to the pure application of the learning methods, we completed our work by using data cleaning and feature selection approaches as well as various interesting techniques for cost sensitive learning.

3.1. Feature Selection. There has been much research effort put into the field of feature extraction.[59] The problem of selecting properties which are responsible for given outputs occurs in various machine learning applications.[60−62] We use feature selection methods with the objective to detect features that are responsible for the underlying class structure. In addition, we search for feature combinations that reflect or even outperform results using all features. We analyze the question of whether unsupervised statistical feature selection methods are able to boost cost sensitive machine learning techniques or not. In this paper, we present results obtained with the method of principal variables.[63] In contrast to principal component analysis (PCA),[64] which is usually applied for data reduction, it does not compute principal components but attempts to assign the optimality property of principal components to single features. PCA suffers from the disadvantage that each principal component is a linear combination of all features, so that data analysis still involves all features. Thus, PCA fails to provide users with interpretable results. It was shown that the method of principal variables gives results comparable to those of PCA with a slightly increased number of selected features.[63] The simplicity of dealing with features instead of noninterpretable linear feature combinations justifies this overhead, especially for the physicochemical features of drugs. Another problem of pure PCA in various software packages is that attributes for the training set are changed, whereas the attributes for the test set remain original. A practical thing to do would be to apply PCA over the whole data set and then divide the set into training and test parts. Since this leads to a test that is not independent from the modeling stage, we avoid this kind of data snooping.

3.2. Support Vector Machine Classification. Support vector machines are powerful data mining methods for classification and regression problems.[35] Their accuracy is excellent, and in many cases they outperform other machine learning methods such as neural networks. SVMs are based on strong theoretical foundations and have their roots in the field of statistical learning which provides the reliable generalization theory.[65] Several properties make this learning method successful, e.g., the kernel trick[66] for nonlinear classification and the sparse structure of the final classification function. In addition, SVMs have an intuitive geometrical interpretation, and a global minimum can be located during the SVM training phase. The concept of support vector machines was introduced by Vladimir Vapnik.[65] The nonlinear SVM hypothesis function is of the form

$$h(x) = \text{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i k_p (x^i, x) + b\right) \forall x \in \mathbb{R}^n \qquad (2)$$

The vector $y \in \{-1; 1\}^l$ reflects the given class labels of the training data, $k_p : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is the kernel function with some parameter $p \in \mathbb{R}$ that has to be chosen by the user. The kernel values can be interpreted as distances between two points in some high-dimensional SVM specific feature space. We refer to ref 66 for a deeper insight into kernel-based methods. The hypothesis (2) is mainly controlled by the so-called Lagrange multipliers α_i ($i = 1, ..., l$) that can be determined via the solution of the L_1-norm or the L_2-norm approach. The difference between these two SVM methods is the way that training errors are treated.[35] The first simply adds the margin errors, whereas the second adds the errors in squared form. It is not fully clear, which method is preferable. The L_1-norm quadratic programming (QP) problem is of the form

$$\min_{\alpha \in \mathbb{R}^l} \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j k_p (x^i, x^j) - \sum_{i=1}^{l} \alpha_i \qquad (3)$$

under the constraints $y^T \alpha = 0$ and $0 \leq \alpha \leq C$. The L_2-norm QP problem is of the form

$$\min_{\alpha \in \mathbb{R}^l} \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j \left(k_p (x^i, x^j) + \frac{\delta_{ij}}{2C_i}\right) - \sum_{i=1}^{l} \alpha_i \qquad (4)$$

under the constraints $y^T \alpha = 0$ and $0 \leq \alpha$. The parameter vector C controls the error penalization. For (3) and (4) unique global solutions do exist. We solve the problem (3) by using the sequential minimal optimization (SMO) algorithm.[67] For the solution of (4), we apply our implementation of the fast SVM training method called nearest point algorithm (NPA).[68]

In section 4 results for both methods will be discussed. The computation of the threshold $b \in \mathbb{R}_+$ is trivial. We refer to ref 35 for a detailed description of the SVM training problem. Due to their special learning mechanism, SVMs are interesting candidates for cost sensitive learning. The decision function (2) only takes examples with positive Lagrange multipliers into account. Due to the SVM opti-

mality conditions,[35] these are the instances close to the boundary, so that the SVM is not affected by a large number of negative instances far away from the hyperplane h.

3.2.1. Weighted SVM Learning Models. The SVM parameter values $C_i \in \mathbb{R}_+$ in (3) and (4) are responsible for the tradeoff between the margin maximization and the error toleration. Often a single value $C_i \equiv C$ is used for simplicity. In ref 69 the authors gave evidence that for unbalanced data sets at least two values should be used to obtain sensitive hyperplanes:[70] $C_i = C^+$ if the ith training point is positive ($y_i = 1$), and $C_i = C^-$ otherwise ($y_i = -1$). In addition to correcting different sizes of the two classes, the (C^+, C^-) model can also capture different costs of false positive and false negative classifications. The combination of the oversampling technique that increases the density of positive instances in order to obtain a well-defined boundary—see section 3.5.—and different error costs that will push the boundary away from the positive class has been proposed in ref 71.

3.2.2. New Kernels for Support Vector Machines. The linear kernel, the Gaussian kernel, also called radial basis function (RBF) kernel, and the polynomial kernel are implemented in various SVM software packages.[72–74] They show promising behavior in different application areas. In our work, we like to introduce two new kernels that are suited for chemical structure classification. In the field of text classification, the family of so-called string kernels[75] has been successfully embedded. For the classification of CYP450 data, we propose the following two distance measures.

The Slater kernel is of the form

$$k_\sigma^S(x^i, x^j) = \exp\left(-\frac{||x^i - x^j||}{2\sigma^2}\right) \quad \sigma \in \mathbb{R}_+ \quad (5)$$

We refer to refs 76 and 77 for a discussion of the Slater (wave) function. We define the Tanimoto kernel which is based on the Tanimoto coefficient[78] as

$$k^T(i, j) = \frac{I_{11}(i, j) + I_{00}(i, j)}{2I_{10}(i, j) + 2I_{01}(i, j) + I_{11}(i, j) + I_{00}(i, j)} \quad (6)$$

The function $I_{a,b}(i, j)$ computes how often $a \in \{0, 1\}$ occurs in x^i, whereas at the same time $b \in \{0, 1\}$ occurs in x^j. This kernel is suitable only for binary data sets. In section 4 we will compare results achieved with the Gaussian kernel and the new kernels (5) and (6).

3.3. Maximum Entropy Classification. Maximum entropy is a powerful method for constructing statistical models for classification tasks. The maximum entropy classification approach is based on the fact that some information about the probability distribution in the training data is known, and thus it is possible to choose a distribution p^* which is consistent and has the highest possible entropy.[79] If $\mathcal{Y}$ denotes the set of classes and $\mathcal{Z}$ denotes the set of possible contexts, the distribution $p^*(y, z)$ should maximize

$$H(p) = -\sum_{(y,z)\in\mathcal{Y}\times\mathcal{Z}} p(y,z) \log p(y,z) \quad (7)$$

H is called the Shannon entropy. The field of information theory was originated by Claude Shannon in 1948. Since then, information theory and Shannon entropy have become

synonymous.[80] H needs to remain consistent with the evidence, which is represented by n binary features f_k ($1 \leq k \leq n$) and their constraints. Therefore, for nondiscrete data of physicochemical properties, we applied a discretization method to build categorical data. The model's expectations are all constrained to match the observed expectations. It can be shown[36] that the solution p^* is of the form

$$p^*(y, z) = \pi \prod_{k=1}^{n} \alpha_k^{f_k} (0 < \alpha_k < \infty) \quad (8)$$

where π is a normalization constant and a is the vector of model parameters. Each a_k represents the weight that is assigned to the feature f_k. The estimation of the feature distribution functions is a convex optimization problem with a unique global maximum. A famous algorithm for solving the optimization problem, i.e., to compute the vector a, is the so-called generalized iterative scaling (GIS) method. GIS is a procedure to find the conditional exponential model weights that define the maximum entropy classifier for a given feature set and training corpus. This procedure is guaranteed to converge on the correct weights. We refer to ref 36 for a detailed description. The *openNLP* maximum entropy package which we have used is freely available from ref 58. It is a mature Java package for training maximum entropy models. Various parameters have to be chosen, e.g., the number of iterations in the GIS algorithm, the smoothing factor to determine how often a feature is shown to the algorithm, and the cutoff to define how often a feature must appear to be considered as relevant.

3.4. Sensitive Quality Management. Learning procedures are controlled by parameters set by the user. In our experiments, we had to tune the kernel parameter value and the error penalization parameters C^+ and C^- for SVM learning as well as the smoothing factor, cutoffs, and the GIS iteration number for ME learning. Finding appropriate values for these parameters is a challenging problem, especially for unbalanced data sets. Tuning the parameters is very important and can be implemented by optimizing a certain quality measure, which is obtained in cross-validation steps. A robust quality measure for classifiers is the Matthews correlation coefficient mcc[16] which is defined as

$$mcc = \frac{tp\cdot tn - fp\cdot fn}{\sqrt{(tp + fn)(tn + fp)(tp + fp)(tn + fn)}} \quad (9)$$

We define tp and tn to be the numbers of the true positive and true negative points. Respectively, fp and fn are the numbers of false positive and false negative classifications. If there is nearly no relationship between the predicted values and the actual values, the correlation coefficient is very low. As the strength of the relationship between the predicted values and actual values increases, so does the correlation coefficient. A perfect fit gives rise to a coefficient of 1.0. Following the ideas in ref 69 we use another effectiveness measure which is defined as

$$E_\beta = 1 - \underbrace{\frac{(\beta^2 + 1)pr \cdot se}{\beta^2 \cdot pr + se}}_{F_\beta} \in [0, 1] \quad (10)$$

and needs to be minimized. It is based on a weighting between sensitivity $se = tp/(tp + fn)$ and precision $pr = tp/(tp + fp)$. One of the advantages of the measure (10) is the parameter $\beta \in \mathbb{R}_+$ that can be adjusted to enforce or diminish the influence of sensitivity. $\beta = 0$ leads to $E_\beta = 1 - pr$, $\beta = \infty$ leads to $E_\beta = 1 - se$. We use $\beta = 1$ which leads to the harmonic mean of sensitivity and precision. F_1 is known as the F-measure, which is used in the field of information retrieval, and therefore eq 10 reduces to $1 - F_1$. Thus, our quality measure is flexible and may be used to tune classifications toward high sensitivity and good overall accuracy. During a cross-validation routine, we compute the numbers of true and false positive points. The quality measure will also be given for the test results.

3.5. Oversampling, Undersampling, and Threshold Moving. In this paper, we address the problem of classifying unbalanced data sets in which negative instances outnumber the positive instances. In addition to the unbalanced class distribution, the costs of classification errors differ in a way where a false negative prediction is much more expensive than a false positive one. Machine learning algorithms usually perform poorly on unbalanced data sets since they comprise a principle that tends to learn a simple hypothesis which classifies all instances negative. Thus, it is necessary to modify either the data or the learning method so that more attention is paid to the positive class. Several approaches have been proposed.[26,71]

As it was described in section 3.2, the penalty for misclassified positive points should be increased to make false negative errors costlier than false positive ones. The implementation of this technique is dependent on the learning method. It was shown that this weighting approach leads to more sensitive results in SVM learning.[69] However, this approach is limited by the learning method to be used and has no influence on the data distribution. We observed that highly unbalanced data sets with a lot of noise force this approach to produce overfitted models.

In this paragraph, we discuss pre- and postprocessing methods for boosting cost sensitive classification. The unbalanced data may be preprocessed by oversampling the minority class or by undersampling the majority class.[71] Although undersampling is a popular method for dealing with the problem of unbalanced data, especially for large data sets where training time of learning algorithms is a bottleneck, valid instances containing valuable information are thrown away. For this reason, we work on the oversampling technique that increases the number of positive examples and does not lead to information loss. This method changes the training data distribution by manifolding higher-cost training points, see ref 81 for an oversampling algorithm.

A popular postprocessing method is the threshold moving technique. The output threshold is moved toward the inexpensive class so that examples of the small class become harder to be misclassified at the cost of some more false positive points. The original or preprocessed data set is used to train some classifiers, and the cost sensitivity is introduced in the test phase. Recent studies had shown that threshold moving is very effective in addressing the problem of cost sensitive learning.[81] For highly unbalanced data sets, where additional false positive points are acceptable when sensitivity increases, threshold moving should be considered. The results achieved by the addition of this technique to machine learning methods will be presented and explained in section 4. In the case of SVM learning, a positive constant is added to the threshold b in (2), so that the function value increases for all test points. For maximum entropy classification, where the output represents the probability for the point to belong to the positive class, we add a constant to the probability value, so that more points pass the class boundary.

3.6. Ensemble Methods. Building ensembles out of multiple classifiers has become an interesting field of research. Ensemble methods are used for the improvement of unstable or weak classifiers. An ensemble expert that combines the outputs from different classifiers is built. Some examples for ensemble techniques are bagging, boosting, and random forests.[31] Usually ensemble techniques are not used for strong classifiers like support vector machines. However, the question arises whether ensemble techniques are able to improve results of such classifiers even more. A study on ensemble methods for SVM has been presented in ref 32. Results of this work are rather pessimistic since neither boosting and bagging nor cluster partitioning methods could improve the performance on several benchmark data sets. A recently published work introduces so-called consensus SVM methods[24] and shows that 81 SVM classification systems are optimal for the data sets used in this analysis. Future work might define other ensemble methods for performance improvement.

Following the general ideas of combining classifiers, we implemented a kernel weighting for SVM classification. For two types of kernels $k^1_{p_1}$ and $k^2_{p_2}$, e.g. the Gaussian and the polynomial kernel,[35] the final kernel function is defined as

$$k_{\gamma, p_1, p_2} = \gamma k^1_{p_1} + (1 - \gamma) k^2_{p_2} \qquad (11)$$

The parameter $\gamma \in [0,1]$ needs to be optimized together with the kernel parameters p_1 and p_2. Such a weighted kernel may improve performance and has the advantage that during parameter optimization either $k^1_{p_1}$ or $k^2_{p_2}$ may be fully rejected by the system. This is the case for $\gamma_{opt} = 1$ or $\gamma_{opt} = 0$. Thus, optimizing[11] helps to find a good kernel family as well. Although our method may not be seen as a real ensemble method, since the weighted kernel leads to a single classifier with only one output per point, the inclusion of two kernels, however, changes the SVM model and may help to increase performance. Such a weighted kernel may be integrated into freely available software packages[72] as well.

4. EXPERIMENTAL RESULTS

In this section, we analyze the classification results of the machine learning methods SVM with different kernels and maximum entropy based on generalized iterative scaling in conjunction with feature selection and several methods for cost sensitive learning. Tests were performed on the ensemble and the binary data sets. In case of the ensemble data set, we have a combination of numerical and categorical descriptors. The model building and prediction were performed through the use of the tab delimited ASCII files. Additionally we reduce the number of features with the unsupervised feature selection algorithm of McCabe.[63] The numbers of selected features are 5, 10, and 20, which are not exclusive choices, but are used to study the possibilities of data reduction. For the oversampling method,[42] we inflated the

HIGHLY UNBALANCED CYP450 DATA OF DRUGS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **97**

Table 1. Characteristics of the 16 Data Sets Which Were Built out of the 2 Basic Data Sets

data set name	1a	1b	1c	1d	1e	1f	1g	1h	2a	2b	2c	2d	2e	2f	2g	2h
# features	5	10	20	557	5	10	20	557	5	10	20	458	5	10	20	458
oversampling		no				yes				no				yes		
# positive training points		35				105				35				105		
# negative training points		150				150				150				150		
# positive test points		13				13				13				13		
# negative test points		65				65				65				65		
basis				ensemble data								binary data				
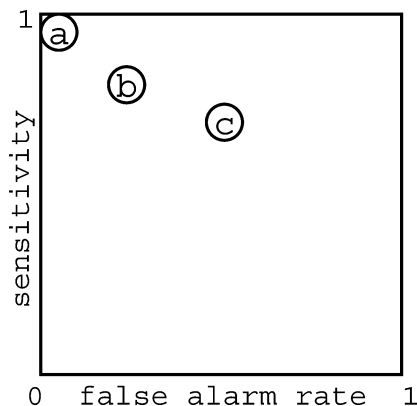
Figure 1. Points in the ROC space.

amount of positives in the training set with a factor of 3 by repeating each feature vector three times. The characteristics of all 16 data sets are given in Table 1.

For parameter optimization we have used a grid search over the parameter space with the quality measure (10) with $\beta = 1$ which was adequate to discriminate between the models of Tables 3−7. The threshold moving technique was applied after the training with a value of 0.5 for SVM learning and 10% for maximum entropy classification. Once an algorithm is trained with the training set, the algorithm classifies the compounds from the test set with feature sets that are new and predicts the 2D6 inhibition. The predicting power for new chemical entities can then be estimated by our quality measure. We have performed a large number of numerical tests, some of which are presented in this section. The generation of the 2D6 model took a few seconds for each run on a PC for the cases of the SVM and ME methods. In our opinion the presentation of classification results is crucial in two manners:

• Usually the performance of a cross-validation run is presented, and for the reader, it is unclear whether the results were obtained with any parameter optimization method or not. Often these test results are highly optimized and not independent.

• The simple presentation of performance in terms of accuracy is insufficient for unbalanced classification problems.

Therefore our experiments are based on the following settings:

• We perform several tenfold cross-validation runs to tune the parameters (grid search). The results are evaluated using (10), and the best parameter tuple is taken for learning the final classifier. The test to be presented is performed on data that were excluded before the optimization process and were not used for the final training as well.

• The presentation of test results is focused on the problem of cost sensitive learning. Therefore we always show sensitivity (hit rate) as well as false positive rate (false alarm rate). These are the characteristics needed to plot results in the so-called receiver operating characteristic (ROC) space,[82] which is a performance graphing method becoming increasingly important for cost sensitive learning. In the ROC space (Figure 1), the point "a" at (0, 1) represents perfect classification. A point in the ROC space is better than another if it is to the northwest of the first. In Figure 1, "b" is better than "c".

4.1. Ensemble Data Set. Based on our unsupervised feature selection method, we selected the top 20 features from our ensemble data set that are summarized in Table 2. All of the selected features fit into the previously described pharmacophor features that are known from pharmacophor modeling and active site studies. They are divided into features that describe the shape and the connectivity of the

Table 2. Selected Features and Their Meaning

chi0	zeroth order connectivity index, sum over all vertices (1/sqrt degree of vertex)
chi0_C	carbon connectivity index (order 0), sum over all carbons (1/sqrt degree of vertex)
TPSA	topological polar surface area
a_nN	number of nitrogen atoms
b_single	number of single bonds
PEOE_VSA_HYD	total hydrophobic vdw surface area based on Gasteiger−Marsili charges
PEOE_PC+	total positive partial charge based on Gasteiger−Marsili charges
MACCS (−60)	number of S=O groups, # [S+] − [O−], [*+*] − [#8−], S
MACCS (165)	number of ring atoms
MACCS (−88)	number of sulfur atoms
MACCS (−56)	number of N bonded to >= 20 and >= 1 C, [#7] (~[#8]) (~[#8]) ~[#6], Q3; ON (O) C
MACCS (−83)	number of heteroatoms in 5 ring, QAAAA@1
Q_VSA_POS	total positive vdw surface area
vsa_acid	VDW acidic surface area (A**2)
SMR	molar refractivity
SMR_VSA0	bin 0 (0.000, 0.110] of molar refractivity
SMR_VSA7	bin 7 (0.560, 10] of molar refractivity
DLI (08)	number of hydroxyl groups, drug like index
kS_sI	Kier atom type E-state sum (sI), [I] [*]
b_triple	number of triple bonds, reactive groups

Table 3. Classification Results for the 8 Ensemble Data Sets Using SVM Classification with a Gaussian Kernel[a]

data set	1a	1b	1c	1d	1e	1f	1g	1h
threshold moving					no			
oversampling			no				yes	
hit rate	0.85	0.77	0.62	0.77	0.69	0.69	0.38	0.69
false alarm rate	0.29	0.20	0.18	0.20	0.20	0.17	0.12	0.17
$1 - F_1$	0.49	0.44	0.52	0.44	0.49	0.45	0.62	0.45
threshold moving					yes			
oversampling			no				yes	
hit rate	1.00	0.77	0.85	0.92	0.92	0.77	0.85	0.85
false alarm rate	0.42	0.22	0.34	0.34	0.23	0.20	0.22	0.23
$1 - F_1$	0.51	0.46	0.52	0.49	0.40	0.44	0.42	0.44

[a] The influence of oversampling is given on the right side of the table.

Table 4. Classification Results for the 8 Ensemble Data Sets Using SVM Classification with Slater Kernel[a]

data set	1a	1b	1c	1d	1e	1f	1g	1h
threshold moving					no			
oversampling			no				yes	
hit rate	0.85	0.62	0.46	0.92	0.69	0.54	0.38	0.62
false alarm rate	0.22	0.15	0.20	0.26	0.14	0.12	0.14	0.12
$1 - F_1$	0.42	0.48	0.63	0.43	0.42	0.50	0.63	0.45
threshold moving					yes			
oversampling			no				yes	
hit rate	0.92	0.92	0.92	0.92	0.92	0.92	0.92	0.92
false alarm rate	0.28	0.23	0.28	0.28	0.26	0.25	0.28	0.29
$1 - F_1$	0.44	0.40	0.44	0.44	0.43	0.41	0.44	0.45

[a] The influence of oversampling is given on the right side of the table.

Table 5. Classification Results for the 8 Ensemble Data Sets Using ME Classification[a]

data set	1a	1b	1c	1d	1e	1f	1g	1h
threshold moving					no			
oversampling			no				yes	
hit rate	0.54	0.54	0.77	0.77	0.54	0.54	0.54	0.92
false alarm rate	0.17	0.17	0.23	0.34	0.17	0.19	0.22	0.29
$1 - F_1$	0.55	0.55	0.47	0.56	0.55	0.56	0.59	0.45
threshold moving					yes			
oversampling			no				yes	
hit rate	0.54	0.54	0.77	0.77	0.77	0.77	0.85	0.92
false alarm rate	0.15	0.15	0.22	0.17	0.15	0.18	0.22	0.29
$1 - F_1$	0.53	0.53	0.46	0.41	0.39	0.43	0.42	0.45

[a] The influence of oversampling is given on the right side of the table.

molecule like the zeroth order connectivity indices chi0 and chi0_C. Then a number of atom counts from the 166-vector MACCS key[37] as well as counts of functional and reactive groups were selected. Finally the SMR, electrotopological state indices for atom types and features, that encode the charge distribution within the molecules, have been selected. Therefore in the data sets with applied feature selection, only 2D features were used to build the models. Comparing the results overall kernels and methods, this did not have an impact onto the final quality of the models. Please note refs 83 (TPSA), 84 (PEOE), 44 (SMR), 40 (DLI(08)), and 41 (b_triple).

In Table 3, we show test results for the ensemble data set. A L_1-norm support vector machine with a Gaussian kernel and different error weights was used. The L_2-norm model showed similar results. Please note the positive influence of threshold moving onto the value of sensitivity. For all test cases, sensitivity increased dramatically. However, this success is adjusted by more false positive points. In addition to this effect, it can be seen that oversampling of the data did not lead to an improvement of sensitivity for the usual SVM, shown at the top of the table. The values of the quality measure degraded. In contrast, for the SVM with threshold moving, the oversampling technique improved the overall test results for all data sets, which can be seen in the last line where the values of our quality measure are given. The best result in terms of the quality measure was achieved for the smallest data set by using threshold moving and oversampling. Thus we conclude that feature selection in combination with cost sensitive learning techniques leads to a sensitive and accurate SVM model.

In Table 4, we show test results for a support vector machine with the Slater kernel. Quality of results is comparable to the Gaussian SVM. Again it can be seen that oversampling could not improve sensitivity of the standard SVM method; the values even decreased. Instead, for some reason specificity increased for all data sets. It turns out that for the Slater kernel the threshold moving technique had high impact onto sensitivity. For all data sets it reached a value of 92% which means that 12 out of 13 positive points had been recognized. Please note that we again have to pay the cost of more false positive points, but, once averaged, we got better quality measure values. An additional oversampling had no influence on the classification results. The classification function remained nearly the same.
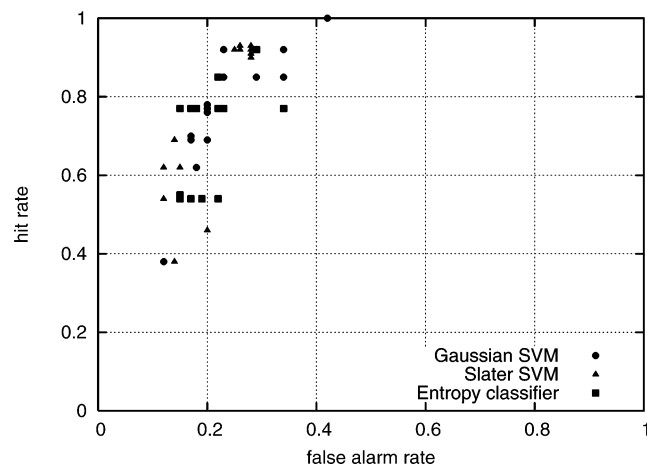
Figure 2. Plot of ROC points for the ensemble data set with attention to the different algorithms.
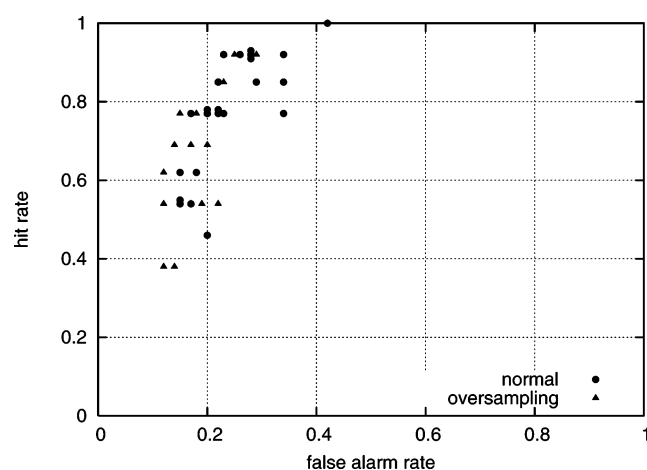

Figure 3. Plot of ROC points for the ensemble data set with attention to the oversampling technique.

The results of our tests with the *openNLP* maximum entropy software package are given in Table 5. With an increasing number of features in the data, sensitivity increased but again at the cost of specificity. In combination with threshold moving, oversampling improved sensitivity. For the smallest data set with five features, sensitivity highly increased without additional costs, which results in a good E_β value. We conclude that cost sensitive learning techniques also work well for maximum entropy classification.

In Figure 2, we compare results of the 3 different classifiers in the ROC space. Obviously the maximum entropy classification results are more stable than the SVM results. Most

of the points are very dense in the ROC space. The SVM methods differ in a way that the Slater kernel led to a majority of points with a high rate of sensitivity with satisfying values of specificity. For both methods it can be seen that higher true negative rates led to dramatically decreased values of the true positive rate. In Figure 3, the influence of oversampling onto the classification results is shown. Oversampling generated classifiers with low false alarm rates, but in contrast to our assumptions, sensitivity decreased.

4.2. Binary Data. In Table 6, we show test results for the binary data set. A L_2-norm support vector machine with our new Tanimoto kernel (6) and different error weights was used. Results for the L_1-norm method are not shown. In our tests, we observed superior behavior of the nearest point method in terms of sensitivity. The main distinctive features of the results are the low true negative rates for small data sets with 5 and 10 features. Since the results for 20 features improved, we conclude that the number of selected features should be larger for the binary data, and feature selection should be applied carefully, if at all. The best results in terms of the quality measure were achieved for the full data set. There, the oversampling technique caused increased values of sensitivity, while, at the same time, the true negative rates slightly decreased.

We compare the results of the support vector machine with the results of the tests with the *openNLP* maximum entropy software package which are given in Table 7. For the two small data sets, the entropy classifier collapsed. All points were declared as negative which is a usual characteristic of classification methods when applied to very unbalanced data sets. They tend to ignore the small class. By using the oversampling method, results improved dramatically and led to good quality measure values. The best results again were obtained for threshold moving in conjunction with oversampling. For the two large data sets, the situation was nearly the same. While recognizing only a small proportion of positive points, the classification of oversampled data sets gave better results for sensitivity. The best quality measure value for all test results presented in this paper was realized by the maximum entropy method for the binary data set with five features and oversampling, as indicated in Table 7. This is mainly due to the very good recognition of negative points (97%).

In Figures 4 and 5, the ROC results are given for the binary data sets. The first distinguishes between the learning methods, whereas the second shows the effect of oversampling onto the classification results. It turns out that the

Table 6. Classification Results for the 8 Binary Data Sets Using SVM Classification with Tanimoto Kernel[a]

data set	2a	2b	2c	2d	2e	2f	2g	2h
threshold moving					no			
oversampling			no				yes	
hit rate	0.77	0.46	0.54	0.62	1.00	0.77	0.54	0.62
false alarm rate	0.69	0.25	0.15	0.12	0.82	0.60	0.15	0.14
$1 - F_1$	0.71	0.66	0.53	0.45	0.67	0.68	0.53	0.47
threshold moving					yes			
oversampling			no				yes	
hit rate	1.00	0.69	0.69	0.85	1.00	0.77	0.69	0.85
false alarm rate	0.82	0.52	0.23	0.28	0.82	0.60	0.23	0.31
$1 - F_1$	0.67	0.68	0.51	0.48	0.67	0.68	0.51	0.50

[a] The influence of oversampling is given on the right side of the table.

Table 7. Classification Results for the 8 Binary Data Sets Using Entropy Classification[a]

data set	2a	2b	2c	2d	2e	2f	2g	2h
threshold moving				no				
oversampling			no			yes		
hit rate	0.00	0.00	0.15	0.38	0.00	0.46	0.38	0.46
false alarm rate	0.00	0.00	0.06	0.08	0.03	0.38	0.18	0.46
$1 - F_1$	1.00	1.00	0.79	0.56	1.00	0.73	0.67	0.48
threshold moving				yes				
oversampling			no			yes		
hit rate	0.00	0.00	0.38	0.54	0.62	0.77	0.54	0.62
false alarm rate	0.00	0.00	0.05	0.06	0.03	0.37	0.22	0.11
$1 - F_1$	1.00	1.00	0.52	0.42	0.30	0.57	0.59	0.43

[a] The influence of oversampling is given on the right side of the table.
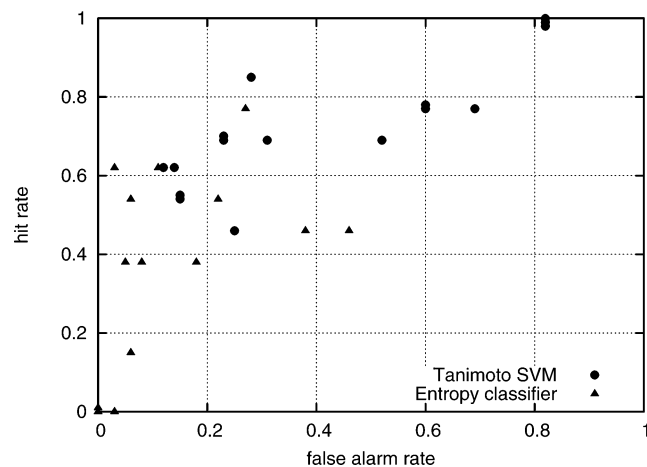
Figure 4. Plot of ROC points for the binary data set with attention to the different algorithms.
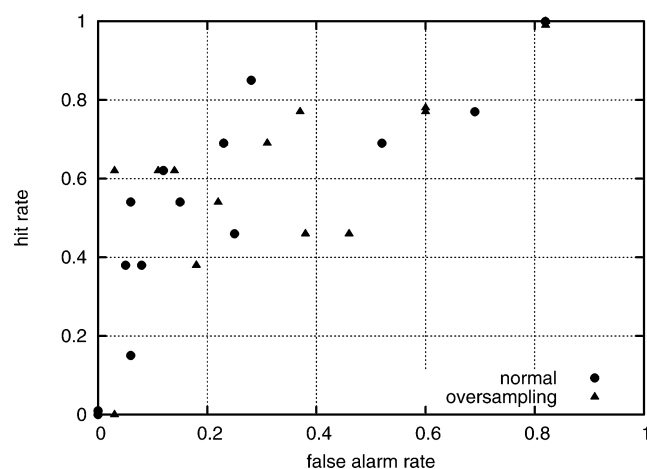
Figure 5. Plot of ROC points for the binary data set with attention to the oversampling technique.

maximum entropy classifier tends to optimize specificity at the cost of the sensitivity values, which are quite poor. The Tanimoto kernel based SVM produced results spread widely in the ROC space. Please note the extreme cases of the Tanimoto kernel SVM at the top right and of the entropy classifier at the bottom left of the ROC area in Figure 4. The results show that the maximum entropy method is conservative, whereas the SVM may be thought of as liberal, since some ROC points are in the upper right-hand side of the space.[82] The comparison of Figures 4 and 5 points out that oversampling slightly improved the sensitivity of the

classifier, but the overall results are not comparable in the way it was possible for the ensemble data set in Figure 3.

Figure 6 gives a representative list of diverse compounds that were included in the training set together with the CYP2D6 inhibition class. It can be seen that even small deviations in the chemical structure like in amitriptyline and imipramine can lead to a different classification. Another example of this kind would be the number of bonds between the aromatic moiety and the nitrogen in venlafaxine and dexfenfluramine that are the same but differ in their CYP2D6 inhibition. In Figure 7, we present selected compounds from our test set on the basis of best E-measures from Tables 3−7 that show some characteristics of the classification algorithms used. Compounds like yohimbine, desmethylsertraline, and prevastatin were classified predominantly correct, whereas others like clemastine, fentanyl, and perazine were often misclassified. This can be understood to some extent by comparing Imipramine from the training set and perazine that was a false negative from the test set. They share some features but differ in an additional sulfur atom for instance. The same is true for the false positives clemastine and fentanyl. Comparing features that were not among the top 20 features listed in Table 2, like the O−N distances from venlafaxine and clemastine in the training set, can lead to a false classification because of a functional moiety. For a number of compounds, the kind of features that were used are important. Herein, the nefazodone classification was enabled using the numerical ensemble features, whereas the binary features lead to misclassifications independent of the method used; conversely for metoprolol where the binary feature set lead to a correct classification. Additionally, we could identify compounds that were better classified by a single algorithm. As an example, we found dextromethorphan correctly classified with the maximum entropy algorithm that shares a common scaffold with codeine from the training set. Most of the other correctly classified compounds can be understood on the basis of their shared similarity. It is clear that only the amount of information that is encoded in the training set will be recognized in the test set independent of the algorithm or kind of feature set.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented our approaches for cost sensitive classification of CYP450 data of drugs by combining machine learning methods with techniques addressing the problem of unbalanced data. We have used support vector machine and maximum entropy classification methods in combination with feature selection as well as oversampling
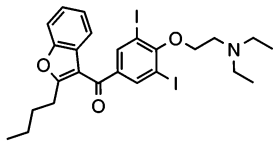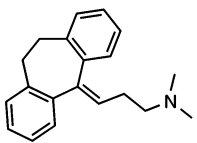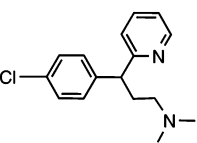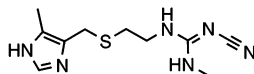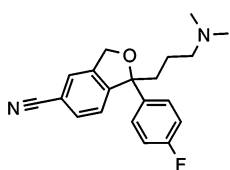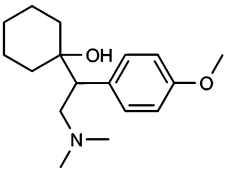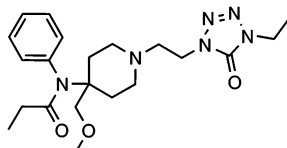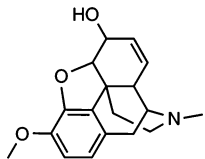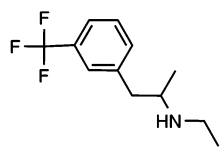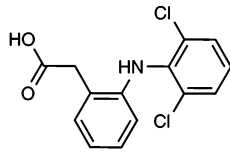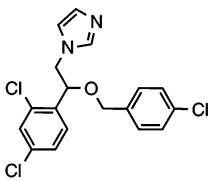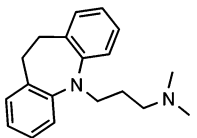
HIGHLY UNBALANCED CYP450 DATA OF DRUGS

J. Chem. Inf. Model., Vol. 47, No. 1, 2007 **101**
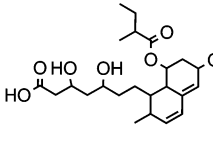

Figure 6. Examples of CYP2D6 inhibition from our training set.

Figure 7. Selected compounds and their classification results on the basis of best E-measures from Tables 3−7.

and threshold moving. We introduced new kernels for SVM classification which we have applied for our tests.

In summary, we have shown that our applied methods are suitable to recognize CYP2D6 inhibition and enable in silico screening of compounds. We have identified the maximum entropy method including discretization of numerical features as equally powerful method in comparison with the support vector machine approach in supervised classification. Herein the use of oversampling and threshold moving in our unbalanced data of CYP2D6 inhibitors was important to yield highly accurate and sensitive classifiers. The unsupervised feature selection method adopted from McCabe selected features that can be understood and interpreted on the molecular level.

We have shown to what extent results can be tuned toward sensitivity, when at the same time overall accuracy is an important factor as well.

Future work will be on the analysis of other members of the CYP450 family which sometimes are even more unbalanced than the current data set. We will further study the problem of unbalanced classification. Particularly, we will improve our SVM kernels, e.g., by embedding a parameter into the Tanimoto kernel. In addition, we will test and improve other quality measures.

Supporting Information Available: List of the compounds from our training and test data sets including smiles codes and names as well as their binary 2D6 classification. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) de Groot, M. J.; Kirton, S. B.; Sutcliffe, M. J. In silico methods for predicting ligand binding determinants of cytochromes P450. *Curr. Top. Med. Chem.* **2004**, *4,* 1803−1824.

(2) Vermeulen, N. P. Prediction of drug metabolism: the case of cytochrome P450 2D6. *Curr. Top. Med. Chem.* **2003**, *3,* 1227−1239.

(3) Lewis, D. F.; Modi, S. Structure-activity relationship for human cytochrome P450 substrates and inhibitors. *Drug Metab. Rev.* **2002**, *34,* 69−82.

(4) Rendic, S.; Carlo, F. J. D. Human cytochrome P450 enzymes: a status report summarizing their reactions, substrates, inducers and inhibitors. *Drug Metab. Rev.* **1997**, *29,* 413−580.

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

(6) Flockhart, D. Cytochrome P450 drug interaction table. http://medicine.iupui.edu/flockhart (accessed Oct 11, 2006).

(7) Kless, A.; Eitrich, T. Cytochrome P450 classification of drugs with support vector machines implementing the nearest point algorithm. *LNAI* **2004**, *3303,* 191−205.

(8) de Graaf, C.; Vermeulen, N. P.; Feenstra, K. A. Cytochrome P450 in silico: an integrative modeling approach. *J. Med. Chem.* **2005**, *48,* 2725− 2755.

(9) van Waterschoot, R. A.; Keizers, P. H.; de Graaf, C.; Vermeulen, N.; Tschirret-Guth, R. A. Topological role of cytochrome P450 2D6 active site residues. *Arch. Biochem. Biophys.* **2006**, *447,* 53−58, 2006.

(10) de Groot, M. J.; Ekins, S. Pharmacophore modeling of cytochromes P450. *Adv. Drug Delivery Rev.* **2002**, *54,* 367−383.

(11) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49,* 2417−2430.

(12) Keizers, P. H.; Schraven, L. H.; de Graaf, C.; Hidestrand, M.; Ingelman- Sundberg, M.; van Dijk, B. R.; Vermeulen, N. P.; Commandeur, J. N. Role of the conserved threonine 309 in mechanism of oxidation by cytochrome P450 2D6. *Biochem. Biophys. Res. Commun.* **2005**, *338,* 1065−1074.

(13) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. Novel approach to predicting P450-mediated drug metabolism: development of a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42,* 1515−1524.

(14) Keizers, P. H.; de Graaf, C.; Kanter, F. J.; Oostenbrink, C.; Feenstra, K. A.; Commandeur, J. N.; Vermeulen, N. P. Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylene-dioxy- N-alkylamphetamines: in silico predictions and experimental validation. *J. Med. Chem.* **2005**, *48,* 6117−6127.

(15) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb. Sci.* **2005**, *24,* 491−502.

(16) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. A support vector machine approach to classify human cytochrome P450 3A4 inhibitors. *J. Comput.-Aided Mol. Des.* **2005**, *19,* 189−201.

(17) Arimoto, R.; Prasad, M.; Gifford, E. M. Development of CYP3A4 inhibition models: comparisons of machine-learning techniques and molecular descriptors. *J. Biomol. Screen.* **2005**, *10,* 197−205.

(18) Kriegl, J. M.; Eriksson, L.; Arnhold, T.; Beck, B.; Johansson, E.; Fox, T. Multivariate modeling of cytochrome P450 3A4 inhibition. *Eur. J. Pharm. Sci.* **2005**, *24,* 451−463.

(19) Singh, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug like molecules. *J. Med. Chem.* **2003**, *46,* 1330−1336.

(20) Kemp, C. A.; Flanagan, J. U.; van Eldik, A. J.; Marechal, J.-D.; Wolf, C. R.; Roberts, G. C. K.; Paine, M. J. I.; Sutcliffe, M. J. Validation of model of cytochrome P450 2D6: an in silico tool for predicting metabolism and inhibition. *J. Med. Chem.* **2004**, *47,* 5340−5346.

(21) Ekins, S.; Berbaum, J.; Harrison, R. K. Generation and validation of rapid computational filters for CYP2D6 and CYP3A4. *Drug Metab. Dispos.* **2003**, *31,* 1077−1080.

(22) Susnow, R. G.; Dixon, S. L. Use of robust classification techniques for the prediction of human cytochrome P450 2D6 inhibition. *J. Chem. Inf. Comput. Sci.* **2003**, *43,* 1308−1315.

(23) O'Brien, S. E.; de Groot, M. J. Greater than the sum of its parts: combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48,* 1287−1291.

(24) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates using support vector machines. *J. Chem. Inf. Model.* **2005**, *45,* 982−992.

(25) Maloof, M. A. *Learning when data sets are imbalanced and when costs are unequal and unknown*; SiteSeer.IST: 2003. citeseer.ist.psu.edu/ maloof03learning.html (accessed Oct 11, 2006).

(26) Barandela, R.; Valdovinos, R. M.; Sanchez, J. S.; Ferri, F. J. The imbalanced training sample problem: under or over sampling? *LNCS* **2004**, *3138,* 806−814.

(27) Japkowicz, N.; Stephen, S. The class imbalance problem: a systematic study. *Intell. Data Anal.* **2002**, *6,* 429−449.

(28) Barandela, R.; Sanchez, J. S.; Garcia, V.; Rangel, E. Strategies for learning in class imbalance problems. *PR* **2003**, *36,* 849−851.

(29) Briem, H.; Günther, J. Classifying kinase inhibitor-likeness by using machine-learning methods. *ChemBioChem* **2005**, *6,* 558−566.

(30) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble methods for classification in cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 1971−1978.

(31) Hall, L. O.; Bowyer, K. W.; Banfield, R. E.; Bhadoria, D. *Comparing pure parallel ensemble creation techniques against bagging*; SiteSeer.IST: 2003. citeseer.ist.psu.edu/hall03comparing.html (accessed Oct 11, 2006).

(32) Dong, Y.-S.; Han, K.-S. *Boosting SVM classifiers by ensemble*; SiteSeer.IST: 2003. citeseer.ist.psu.edu/727102.html (accessed Oct 11, 2006).

(33) Schapire, R. E. *A brief introduction to boosting*; SiteSeer.IST: 1999. citeseer.ist.psu.edu/schapire99brief.html (accessed Oct 11, 2006).

(34) Abe, S. *Support vector machines for pattern recognition;* Springer: 2005.

(35) Cristianini, N.; Shawe-Taylor, J. *An introduction to support vector machines and other kernel-based learning methods;* Cambridge University Press: Cambridge, U.K., 2000.

(36) Ratnaparkhi, A. *A simple introduction to maximum entropy models for natural language processing*; SiteSeer.IST: 1997. citeseer.ist.psu.edu/ 128751.html (accessed Oct 11, 2006).

(37) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. MACCS: reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42,* 1273−1280.

(38) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching *J. Chem. Inf. Comput. Sci.* **1998**, *38,* 983−996.

(39) Potter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41,* 478−488.

(40) Xu, J.; Stevenson, J. A new approach to measure drug-like compounds and their diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40,* 1177−1187.

(41) Oprea, T. I. Property distribution of drug-related chemical databases. *J. Comput.-Aided. Mol. Des.* **2000**, *14,* 251−264.

(42) Rajarshi, G.; Jurs, P. C. Determining the validity of a QSAR model - a classification approach. *J. Chem. Inf. Model.* **2005**, *45,* 65−73.

(43) MOE (The Molecular Operating Environment) Version 2005.06, Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7. http://www.chemcomp.com (accessed Oct 11, 2006).

(44) Wildman, S. A.; Crippen, G. M. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39,* 868− 873.

(45) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property relations. In *Reviews of Computational Chemistry;* Boyd, D., Lipkowitz, K., Eds.; VCH Publishers: Inc., 1991; pp 367−422.

(46) Hall, L. H.; Kier, L. B. The nature of structure-activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.* **1977**, *12,* 307−314.

(47) Randic, M. On molecular identification numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24,* 164−175.

(48) 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.

(49) Kier, L. B.; Hall, L. H. *Molecular structure description: the electrotopological state;* Academic Press: San Diego, CA, 1999.

(50) Randic, M. Graph theoretical approach to local and overall aromaticity of benzenoid hydrocarbons. *Tetrahedron* **1975**, *31,* 1477−1481.

(51) Schuur, J. H.; Setzer, P.; Gasteiger, J. The coding of the three dimensional structure of molecules by molecular transforms and its

application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36,* 334−344.

(52) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3,* 537−547.

(53) Gasteiger, J. Empirical methods for the calculation of physicochemical data of organic compounds. In *Physical Property Prediction in Organic Chemistry;* Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer: Heidelberg, 1988; pp 119−138.

(54) Ihlenfeldt, W. D.; Gasteiger, J. All descriptors for ensembles and molecules. *J. Comput. Chem.* **1994**, *8,* 793−813.

(55) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34,* 109−116.

(56) Thrun, S.; Mitchell, T. M. *Learning one more thing*; SiteSeer.IST: 1995. http://citeseer.ist.psu.edu/141692.html (accessed Oct 11, 2006).

(57) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: data mining, inference and prediction;* Springer: 2001.

(58) Baldridge, J.; Bierner, G.; Friedmann, E.; Morton, T. The openNLP maximum entropy package for classification, 2006. https://sourceforge.net/projects/maxent (accessed Oct 11, 2006).

(59) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused compound collections. *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 993−999.

(60) Wegner, J. K.; Froehlich, H.; Zell, A. Feature selection for descriptor based classification models (1. theory and GA-SEC algorithm). *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 921−930.

(61) Wegner, J. K.; Froehlich, H.; Zell, A. Feature selection for descriptor based classification models (2. human intestinal absorption HIA). *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 931−939.

(62) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44,* 1630−1638.

(63) McCabe, G. P. Principal variables. *Technometrics* **1984**, *26,* 137−144.

(64) Jolliffe, I. T. *Principal component analysis;* Springer: New York, 1986.

(65) Vapnik, V. N. *Statistical learning theory;* John Wiley & Sons: New York, 1998.

(66) Schölkopf, B. *The kernel trick for distances*; SiteSeer.IST: 2000. citeseer.ist.psu.edu/543420.html (accessed Oct 11, 2006).

(67) Platt, J. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods−Support Vector Learning;* Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT Press: Cambridge, MA, 1999; pp 185−208.

(68) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. A fast iterative nearest point algorithm for support vector machine classifier design. *IEEE Trans. Neur. Net.* **2000**, *11,* 124−136.

(69) Eitrich, T.; Lang, B. Efficient optimization of support vector machine learning parameters for unbalanced datasets. *J. Comput. Appl. Math.* **2006**, *196,* 425−436.

(70) Drish, J. *Obtaining calibrated probability estimates from support vector machines*; SiteSeer.IST: 2001. citeseer.ist.psu.edu/drish01obtaining.html (accessed Oct 11, 2006).

(71) Akbani, R.; Kwek, S.; Japkowicz, N. Applying support vector machines to imbalanced datasets. *LNCS* **2004**, *3201,* 39−50.

(72) Chang, C. C.; Lin, C. J. *LIBSVM: a library for support vector machines;* Department of Computer Science and Information Engineering, National Taiwan University: Taipei, Taiwan, 2006

(73) Joachims, T. *SVM-light support vector machine;* Department of Computer Science, Cornell University: Ithaca, NY, 2004.

(74) Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques;* Morgan Kaufmann: San Francisco, CA, 2005.

(75) Lohdi, H.; Saunders, C.; Shawe-Taylor, C. J.; Cristianini, N.; Watkins, C. Text classification using string kernels. *JMLR* **2002**, *2,* 419−444.

(76) Slater, J. C. Atomic shielding constants. *Phys. Rev.* **1930**, *36,* 57−64.

(77) Slater, J. C. Analytic atomic wave functions. *Phys. Rev.* **1932**, *42,* 33−43.

(78) Swamidass, S. J.; Chen, J.; Bruand, J.; Phung, P.; Ralaivola, L.; Baldi, P. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics* **2005**, *21,* 359−368.

(79) Nallapati, R. *Discriminative models for information retrieval*; Site-Seer.IST: 2004. citeseer.ist.psu.edu/654337.html (accessed Oct 11, 2006).

(80) Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27,* 379−423, 623−656.

(81) Zhou, Z.-H.; Liu, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Knowl. Data Eng.* **2006**, *18,* 63−77.

(82) Fawcett, T. *ROC graphs: notes and practical considerations for researchers*; SiteSeer.IST: 2004. citeseer.ist.psu.edu/646695.html (accessed Oct 11, 2006).

(83) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* **2000**, *43,* 3714− 3717.

(84) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges. *Tetrahedron* **1980**, *36,* 3219−3228.