Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features

Markus Wagener* and Vincent J. van Geerestein

Department of Molecular Design and Informatics, N. V. Organon, P. O. Box 20,
5340 BH Oss, The Netherlands

Received August 2, 1999

Using decision trees, a model to discriminate between potential drugs and nondrugs has been developed. Compounds from the Available Chemical Directory and the World Drug Index databases were used as training set; the molecular structures were represented using extended atom types. The error rate on an independent validation data set is 17.4%. The number of false negatives can be reduced by penalizing the misclassification of drugs so that 92 out of 100 potential drugs are correctly recognized. At the same time, 34 out of 100 nondrugs are classified as potential drugs. The predictions of the model can be used to guide the purchase or selection of compounds for biological screening or the design of combinatorial libraries. The visualization of the generated models in the form of colored trees allowed us to identify a few, surprisingly simple features that explain the most significant differences between drugs and nondrugs in the training set: Just by testing the presence of hydroxyl, tertiary or secondary amino, carboxyl, phenol, or enol groups, already three quarters of all drugs could be correctly recognized. The nondrugs, on the other hand, are characterized by their aromatic nature with a low content of functional groups besides halogens. The general applicability of the model is shown by the predictions made for several Organon databases.

INTRODUCTION

The advent of high-throughput screening has triggered the creation of extensive screening efforts in the pharmaceutical industry. The need for more and more compounds that can be screened is fulfilled by compound acquisition from outside sources or by combinatorial chemistry approaches.[1−3] Traditionally, the major goal when selecting the compounds to buy or synthesize has been to increase the chemical diversity of the structures available for screening.[4−7] Little or no attention has been paid to other properties that finally influence whether a compound will become a successful drug or not, such as toxicity, bioavailibility, ease of synthesis, etc. This neglect has even lead to a shift in the profile of compound libraries into an unfavorable direction since the introduction of combinatorial chemistry.[8]

There are two ways to take into account additional properties that are important for a compound to exhibit druglike behavior. One possibility is to explicitly predict each of these properties for the structures under consideration and use the results as additional selection criteria. Although this approach mimics the normal experimental procedure in pharmaceutical research, it is complicated by the need for a predictive model for each of the relevant parameters.

A second, more pragmatic approach is to directly predict whether a compound is a potential drug without explicitly using decision criteria like toxicity, bioavailibility, ease of synthesis, etc.: If a compound is structurally more similar to examples of known drugs than to nondrugs, it can be assumed that it might indeed be a potential drug. In this way, all biological, chemical, or economic reasons for a compound being druglike are implicitly taken into account and need not be separately modeled.

Recently, methods have been reported in the literature that use the latter approach to predict whether a compound might be a drug or not. They are based either on neural networks[9,10] or on optimization of a linear scoring function with a genetic algorithm.[11] For all three methods good results in distinguishing examples of known drugs from nondrugs are reported.[12] However, none of the methods provide a simple analysis of the structural features that are important for achieving this distinction and which can be used as guidelines for medicinal chemists. Only Ajay et al. give a list of 78 out of the 173 descriptors used that were considered to be important for the classification.

In the following we present a method for distinguishing between potential drugs and nondrugs that is based on decision trees.[13] While offering the same accuracy of predictions as the methods mentioned above, our approach additionally provides comprehensible models that can easily be visualized and inspected. On the basis of this visualization it is straightforward to identify important structural features and to explain single predictions.

MATERIALS AND METHODS

Data Sets. In order to generate a model that can predict whether a molecular structure represents a potential drug or not, examples of both, compounds known to be drugs and compounds known to be nondrugs, are needed. While the former is readily available in the form of databases containing marketed and development drugs, it is much harder to come up with the latter—compounds that are known to be non-

* To whom correspondence should be addressed. Phone: +31 412 661380. Fax: +31 412 662539. E-mail: M.Wagener@Organon.Oss.AkzoNobel.NL.

10.1021/ci990266t CCC: $19.00 © 2000 American Chemical Society
Published on Web 02/25/2000

drugs. We chose to take our training examples from the World Drug Index (WDI)[14] and the Available Chemical Directory (ACD).[15] We regarded the compounds from the former as drugs, whereas the compounds from the latter were presumed to be nondrugs. Although many structures in the ACD are known to be drugs or have druglike structures, the ACD represents a large collection of chemicals with a low structural bias toward special classes of compounds (e.g. drugs, agrochemicals, dyestuffs, etc.). Therefore, it is reasonable to assume that the ACD contains fewer druglike compounds than a database dedicated to marketed and development drugs.

The structures of the two databases were preprocessed in a way similar to the one described in ref 9:

(1) Reactive or otherwise unsuitable compounds were removed (e.g. acid halides, anhydrides, metal containing compounds, compounds with a molecular weight below 150 or above 1000, etc.).

(2) Counterions and solvent molecules were removed.

(3) Where possible, charged acidic or basic groups were neutralized by adding or removing protons.

(4) Duplicates within each individual database were removed.

(5) Compounds shared by both databases were removed from the ACD.

The second and third steps converted the structures into a standardized form, allowing the removal of duplicates within and between the two databases. Since all compounds in the WDI are considered to be drugs, duplicates between the two databases were removed only from the ACD. After the preprocessing steps, 38 416 WDI compounds and 169 331 ACD compounds remained, a reduction by 29.5 and 23.9%, respectively.

The preprocessed databases served as a basis for three different data sets necessary to develop and validate a model that can distinguish between those compounds that are druglike and those that are not:

(1) A training data set of 10 000 compounds, with 5000 compounds randomly drawn from each of the two databases, respectively. This data set was used to generate the models.

(2) A test data set of 20 000 compounds, with 10 000 compounds randomly drawn from the ACD and the WDI, respectively, after the compounds in the training set have been removed. On the basis of this data set, alternative models were compared.

(3) A validation data set of 177 747 compounds containing all the compounds from the ACD (154 331 compounds) and the WDI (23 416 compounds) that have not been part of the training or test data sets. The validation data set was used to assess the prediction error of the model that was finally chosen. Since the number of compounds from the ACD and the WDI differ in this data set, two separate prediction errors were calculated for the WDI and the ACD compounds, respectively, and then the mean of these two error values was taken.

Descriptors. For the intended analysis, the compounds in the two databases have to be represented in such a way that features relevant for the distinction between drugs and nondrugs are taken into account. Additionally, the descriptors should be broadly applicable, as compounds in the ACD and WDI represent a highly diverse set of structures. We chose to use the descriptors that Ghose and Crippen developed for the prediction of octanol/water partition coefficients,[16] a set of descriptors that has also been used in ref 9. We consider these descriptors especially suited for our analysis, as they were originally intended for prediction of log *P* values, a parameter important for assessing the usefulness of potential drugs.

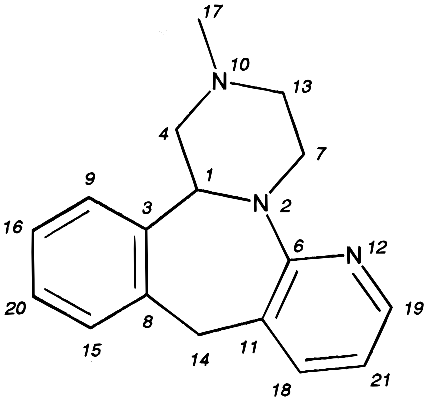


Figure 1. Structure of Mirtazapine. The atom types of all non-hydrogen atoms according to the Ghose−Crippen scheme are listed in Table 1.

Table 1. Atom Types of All Non-Hydrogen Atoms in Mirtazapine (cf. Figure 1)

atom type	atom type no.	atom label
carbon in CH_2R_2	2	14
carbon in CH_3X	5	17
carbon in CH_2RX	6	4, 7, 13
carbon in CHR_2X	8	1
aromatic carbon in RC(H)R	24	9, 15, 16, 18, 20, 21
aromatic carbon in RC(R)R	25	3, 8, 11
aromatic carbon in RC(H)X	27	19
aromatic carbon in RC(X)X	29	6
nitrogen in $N(alkyl)_3$	68	10
nitrogen in aryl-$N(alkyl)_2$	71	2
nitrogen in pyridine-like aromates	75	12

These descriptors characterize each structure by the count of 120 different atom types. The atom types not only take into account atomic number and hybridization of the atom itself but also the nature of the atoms directly adjacent. For nitrogen, e.g., there are 13 different types, depending on the adjacent groups: nitrogen in an aliphatic primary amine, nitrogen in an amide, nitrogen in a nitro group connected to an aliphatic rest, etc. Some atoms (e.g. quaternary nitrogen atoms, nitrogen with one heteroatom and two carbon atoms adjacent, sulfur in polysulfides, etc.) do not correspond to any of the original 120 atom types devised by Ghose and Crippen. For these atoms (0.37% of the 8 798 199 atoms in the combined ACD and WDI data sets) an additional atom type was introduced. The atom types found in the structure shown in Figure 1 are listed in Table 1; a complete list of all descriptors used can be found in the Supporting Information.

Decision Trees. Decision trees are used to solve classification problems, e.g. assigning a chemical structure of unknown class to one of several structure classes.[13,17] The name *decision tree* is due to the fact that the classification is done using a set of tests (or decisions) that are arranged in the form of a tree. Figure 2 shows a very simple decision tree that can be used to classify some oxygen-containing structures. The decision tree consists of a set of nodes; the

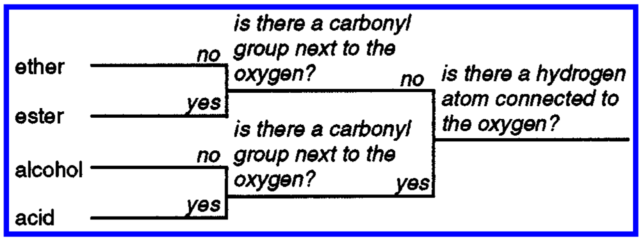

ether
ester
alcohol
acid
no
yes
no
yes
is there a carbonyl group next to the oxygen?
is there a carbonyl group next to the oxygen?
no
yes
is there a hydrogen atom connected to the oxygen?

Figure 2. Simple decision tree to classify some oxygen-containing compounds.

first of them is the so-called root node (see the right-hand side of Figure 2). At each of the nodes one feature of the structure to be classified is tested. If the outcome of the test is true, the lower branch of the tree is followed; otherwise the classification continues with the upper branch. This process is finished once a leaf node, i.e. a node without branches and thus without further tests, is encountered. Here, the class of the leaf node is assigned to the unknown compound. So, a compound with a hydrogen atom attached to the oxygen atom, but no carbonyl group next to it, will be classified as an alcohol. This example shows that each path from the root of the tree to a leaf node corresponds to a rule to identify a class. Therefore, the tree shown in Figure 2 contains four rules, each one of them identifying ethers, esters, alcohols, and acids, respectively.

Classifiers like decision trees are normally induced or trained with a data set of known classifications. The training of a decision tree proceeds as follows: First, a single feature or descriptor is identified that splits the entire training data set into two more homogeneous subsets. This is achieved by evaluating all potential partitionings of the training data set and choosing the one that enriches the known classes in the generated subsets the most. One way to quantify the enrichement obtained with a given partitioning is to compare the entropy of the class distribution before and after the split.[13] In the following steps, the resulting subsets are again split into sub-subsets, generally using different descriptors. This procedure continues until no further significant splits can be found. The result is often a very complex tree that is fitted too closely to the training data. Therefore, those parts that give rise to a high estimated prediction error are again removed from the tree. Several methods are in use to estimate the prediction error, e.g. cross-validation schemes or estimations based on the training set error.[13]

The accuracy of classifiers can often be further improved by generating several classifiers for the same problem and combining their predictions in a special voting procedure. A method that has been successfully applied to combine several decision trees is called *boosting*:[18] During the training of boosted decision trees a weight is maintained for each data point in the training set reflecting its importance. Adjusting these weights emphasizes different parts of the training data set. The first step of boosting involves the generation of a tree using the procedure outlined in the previous paragraph with equal weight assigned to each data point. Then, the weights of those data points are increased that could not correctly be classified by the first tree. On the basis of the updated weights a second tree is generated that consequently focuses on the data points misclassified by the previous tree. This process continues until the desired number of decision trees is generated that will be “experts” for certain classes of data points (compounds). Finally, when a prediction is made using a set of boosted trees, the votes are combined into a prediction so that these “expert trees” get a higher weight for the classes of compounds they can predict the best.

Decision trees combine a number of advantageous features of other popular classifiers, like neural networks or linear discriminant analysis: First, their training is so fast that they can be applied to very large data sets, second, the resulting decision trees can often be interpreted, third, they can deal to some extent with nonlinearities in the data, and, fourth, irrelevant descriptors are to some extent ignored during the generation of the tree.

In our study, the program package C5.0 was used to train and interpret all decision trees.[19] The training of decision trees with C5.0 can be influenced with three parameters; the first two of them are influencing the size of the trees generated, and the third determines the number of trees used for boosting: (1) the minimum leaf size *M* controlling the minimum number of training cases that have to be mapped onto a leaf node in the decision tree; (2) the pruning confidence *CF* controlling the estimated maximum number of misclassifications that are allowed per leaf;[20] (3) the number of boosted trees *NB* that are generated.

RESULTS

Selecting the Most Predictive Model. When training a classifier such as a decision tree, it is possible to control the complexity of the model to be generated. This complexity can be measured in different ways, e.g. the number of descriptors that are used in the model, or, in the case of decision trees, the number of leaf nodes in the tree. The complexity of the model largely influences the quality of the model: The more complex it is, the better the properties of a particular training data set are accounted for, and, consequently, the lower is the error when applying the model to the training data.

However, the real goal is not to predict the classes of compounds in the training set but to apply the decision tree to new compounds, whose classes are not yet known. To assess the predictive power of the model on those unseen cases, an independent test set can be used. The influence of the model complexity on the prediction of the test set, i.e., on the predictive power, is different now: If the model is too simple, it does not contain all the important features of the underlying phenomenon. On the other hand, if the model is too complex, it is completely fitted to the particular features of the training data set, i.e., the model is *overtrained*. In both cases, the optimum model complexity has not been found and the predictions made using the model could still be improved.

In order to find a model with optimum predictive power to discriminate between drugs and nondrugs, the relationship between the prediction error and the tree size, respectively, and the parameters *CF* and *M* controlling the training with C5.0 has been studied. The two parameters have been changed systematically, and for each combination of the parameters a decision tree has been generated. The training of each of the trees was based on the same training set of 10 000 compounds, and the prediction error was always tested using the same test set of 20 000 compounds. The results of these calculations are shown in Figure 3.

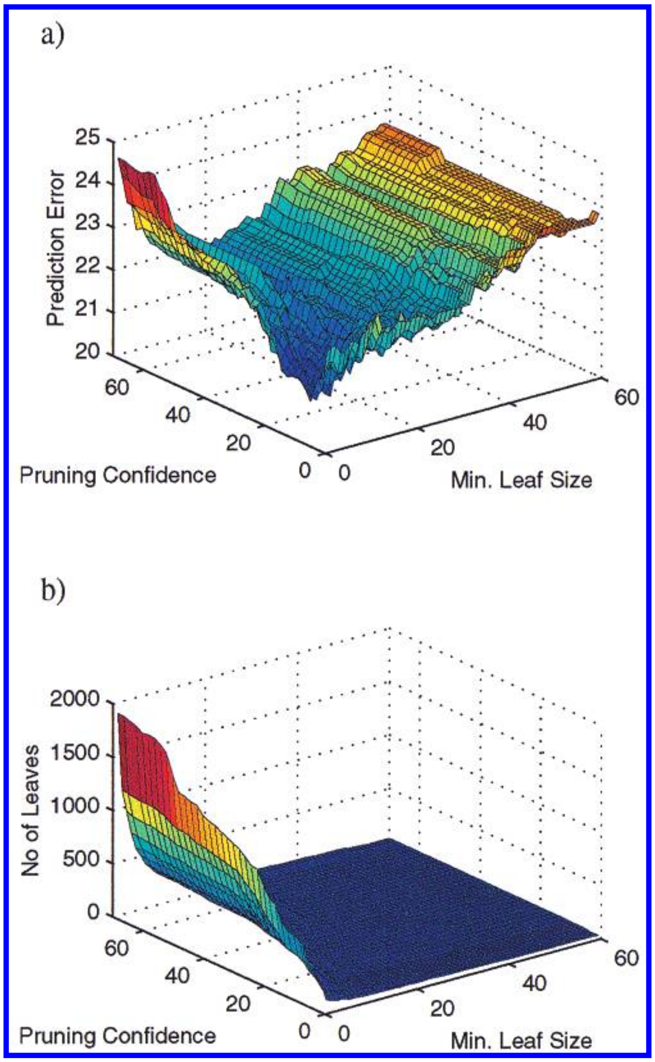


Figure 3. Relationship between the training parameters of the decision tree and (a) the prediction error and (b) the size of the tree. The parameters controlling the training are the pruning confidence *CF* and the minimum leaf size *M*, the prediction error is given as the percentage of incorrect predictions, and the tree size as the number of leaf nodes in the tree.

Figure 3a shows the relationship between the error rate (i.e. the number of incorrect predictions expressed in percent) and the training parameters of C5.0. Since the minimum and maximum prediction errors differ by only 3.7% (24.6%−20.9%), the prediction error is rather stable with respect to the training parameters. Nevertheless, a distinct area of low prediction errors, which is colored in dark blue, can be identified in the plot.

In contrast to the prediction errors, the sizes of the corresponding trees (Figure 3b) differ to a large extent: The smallest tree generated has only 19 leaf nodes, whereas the largest tree has more than 1900 leaf nodes. The comparison of the two parts of Figure 3 reveals that the blue area of comparatively low error in Figure 3a corresponds to medium-sized trees in Figure 3b. Therefore, trees with more than 300 leaf nodes are likely to be overtrained, whereas trees with fewer than 100 leaves cannot describe all the effects present in the data sets.

The above analysis leads to several conclusions: To obtain a tree with maximum predictive power, it should be trained with a low value for the minimum leaf size *M* and a moderately low value for the pruning confidence *CF*. The tree with the lowest prediction error of 20.9% in the above experiments was obtained with $M = 2$ and $CF = 8$. Although this is the best tree found for predicting whether a compound is druglike or not, its size of 266 leaf nodes prevents an easy inspection and interpretation. However, the prediction error does not change very much with the tree size. Even a much smaller tree with only 23 leaf nodes ($M = 56$, $CF = 1$) has a prediction error of only 23.6%, i.e., only 2.7% worse than the best tree obtained. Inspection of the upper levels of the two trees confirms that the general structure of the trees is quite similar. The differences are mostly in the deeper levels of the trees where the larger tree splits the training data into finer groups and thereby reduces the prediction error by another 2.7%. Consequently, smaller trees generated with different training parameters can safely be used to inspect the main structure of the classifiers generated.

The prediction error of the two trees mentioned above on a validation set that was not used for model selection is 21.4% for the tree with 266 leaf nodes and 23.3% for the tree with 23 leaf nodes.

Interpretation of the Model. To support the analysis of decision trees, we have developed an interactive tree browser that visualizes the structure of a tree together with information on the training data such as error rates. Different views of the tree can be generated, and by interactively selecting one or several nodes in the tree it is possible to display the corresponding chemical structures.

Figure 4 shows such a visualization of the tree with 23 leaf nodes and a prediction error of 23.6% on the test set that has been discussed in the previous section. The leaves of the tree are labeled with either D or N, corresponding to the two classes *drugs* and *nondrugs*. Behind the label two numbers are given: First, the number of compounds from the training set that are mapped onto this leaf node, and second, the number of misclassified compounds in that leaf node, i.e., the numbers of Ds in a node labeled N, and vice versa.

At each nonterminal node of the tree, the condition of the split at that node is given. If the indicated condition is met, the lower branch is used; otherwise the upper branch is taken. The first split, e.g., is testing whether there are more than 0 alcohol−oxygen atoms in the compound to be classified. If this happens to be the case, the compound is immediately classified as a potential drug. Otherwise, if there is no alcohol−oxygen atom in the compound, the classification is not yet finished and continues with a test on tertiary aliphatic amines.

Using this tree, the structure of Mirtazapine from Figure 1 is classified as follows: Mirtazapine has no alcohol oxygen (type O:alcohol), but one tertiary aliphatic amine (atom 10 in Figure 1, type N:Al_3N), and three methylene carbons that are bonded to two hydrogen atoms, one carbon atom, and one heteroatom (atoms 4, 7, and 13 in Figure 1, type C:CH_2RX). Therefore, as the number of alcohol oxygens in Mirtazapine is zero, the upper branch is taken at the first node of the tree shown in Figure 4. Then, since there is one tertiary aliphatic amine, the lower branch is taken at the next node. Finally, since there are three methylene carbons of type C:CH_2RX in Mirtazapine, leaf node 22 is reached. Consequently, since this leaf node carries the label D, Mirtazapine is classified as a drug.

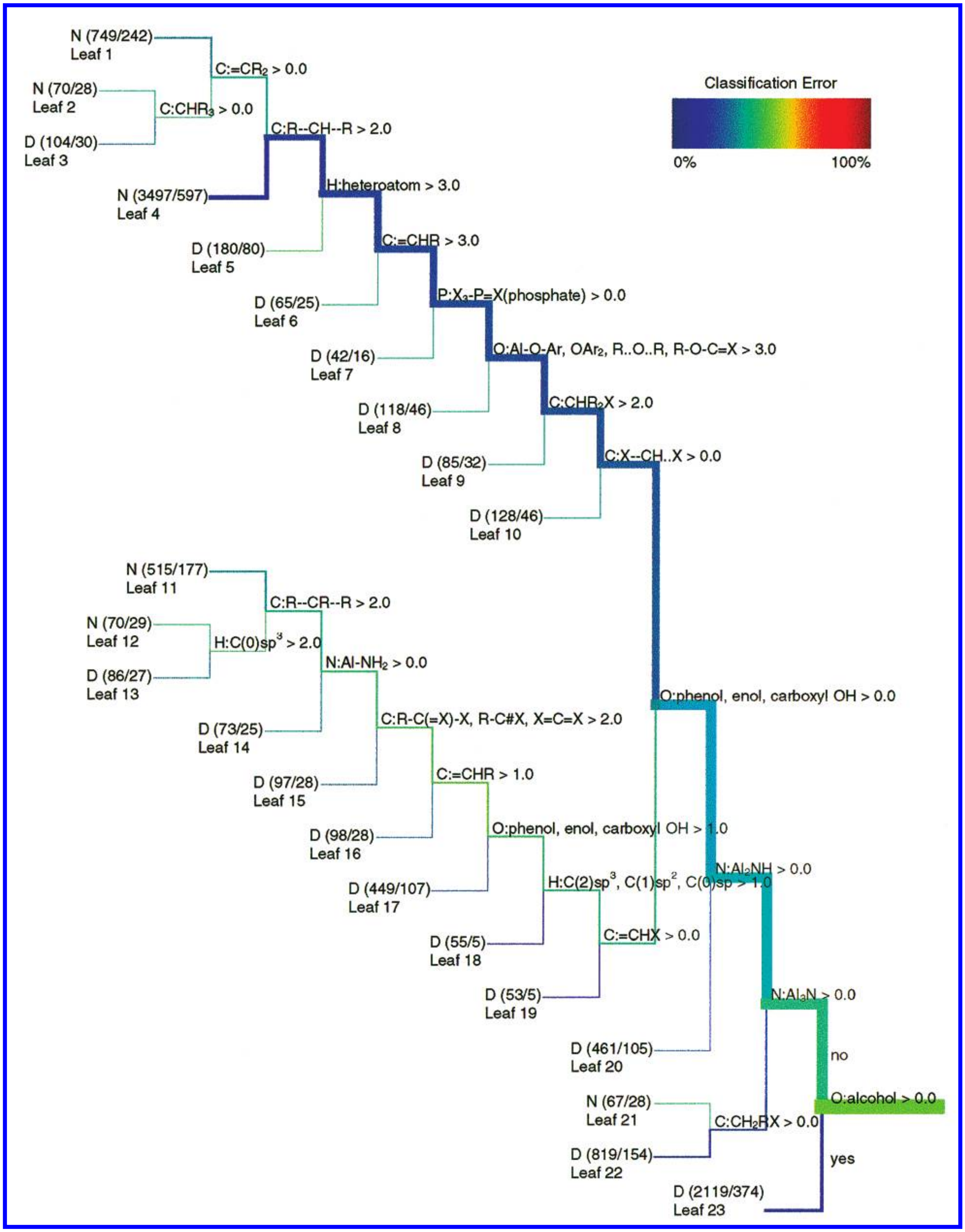


Figure 4. Simplified decision tree to predict whether a compound is likely to act as a drug or not. Each leaf of the tree is labeled either with D for drugs or with N for nondrugs. Behind the label the number of compounds in that leaf and the number of classification errors are given. The line width indicates the number of compounds; the classification error of the training data set is color-coded (blue = 0%, green = 50%, and red = 100% error). The descriptors and tests used for the classification are given above each nonterminal node. The following abbreviations are used: R represents any group linked through a carbon; X represents any heteroatom; Al and Ar represent aliphatic and aromatic groups, respectively; −, =, and # represent single, double, and triple bonds, respectively; - - represents aromatic bonds as in benzene or delocalized bonds such as in nitro groups; •• represents aromatic bonds such as the CN bond in pyrrole; $C(n)sp^x$ represents a carbon with formal oxidation number n and hybridization sp^x.

The width of the lines in Figure 4 is proportional to the number of compounds from the training set that are mapped onto this part of the tree. Accordingly, the width of the root at the rightmost part of Figure 4 corresponds to the complete training set, i.e., 10 000 compounds, the lower branch corresponds to 2119 compounds, and the upper branch to the remaining 7881 compounds.

The error rate of each node based on the training set is color-coded, with blue corresponding to 0%, green to 50%, and red to 100% error, respectively. Therefore, the root of the tree is colored in green, as the complete training set consists of 5000 compounds from the WDI and 5000 compounds from the ACD, corresponding to an error rate of exactly 50% at this node. The color of the nodes gradually changes from green to blue, as the classification error decreases while going from the root to the leaves.

The tree in Figure 4 essentially consists of two main branches with 12 and 13 levels of nodes, respectively. Many small leaf nodes labeled as drugs are split off directly from these two branches. The nondrug compounds are, with one exception, all classified by leaves in the lowest levels of the tree. The most prominent one of them, leaf 4 in Figure 4 contains $3497-597 = 2900$ nondrug compounds, more than half of all nondrugs in the training set. Altogether, there are 17 leaves labeled as drugs, but only 6 leaves labeled as nondrugs.

The structure of the tree reveals a surprisingly simple scheme to distinguish between potential drugs and nondrugs: The presence of some elementary features, most often functional groups, is tested one after the other. If these features are present, the compound is immediately considered as a potential drug. The very first split in the tree already illustrates this striking behavior: If the compound under consideration contains a hydroxy group at all, it is classified as a drug without any further tests. The features most important for a compound to be considered as a drug are functional groups containing either oxygen or nitrogen, notably alcohols, tertiary and secondary aliphatic amines, and phenols, enols, and carboxyl groups, i.e. a hydroxy group at a sp^2-carbon atom.

Interestingly, the tree identifies many features important for druglike behavior, but almost no clue is given on properties that would render a compound useless as a drug. Rather, nondrugs are the featureless compounds remaining after all the compounds containing the essential functional groups have been removed.

To give an impression of the types of structures that are mapped onto the different leaves of the tree in Figure 4, two representative structures for each leaf node (one from the ACD and one from the WDI, respectively) are shown in Figure 5. These representative structures were selected on the basis of their BCI fingerprints.[21] The structures are labeled with the number of the leaf they belong to; leaf 1 is the one in the upper left corner of Figure 4.

It can be seen that the leaves in the first levels of the tree can contain rather diverse compounds (cf. the two example structures for leaf 23 or for leaf 21). The farther away from the root the leaves are, the less diverse the corresponding compounds are (cf. leaf 16 with two brominated cinnamic acids). This is clearly the result of the increasing number of constraints that the different decision nodes impose on the compounds the farther away they are from the tree root.

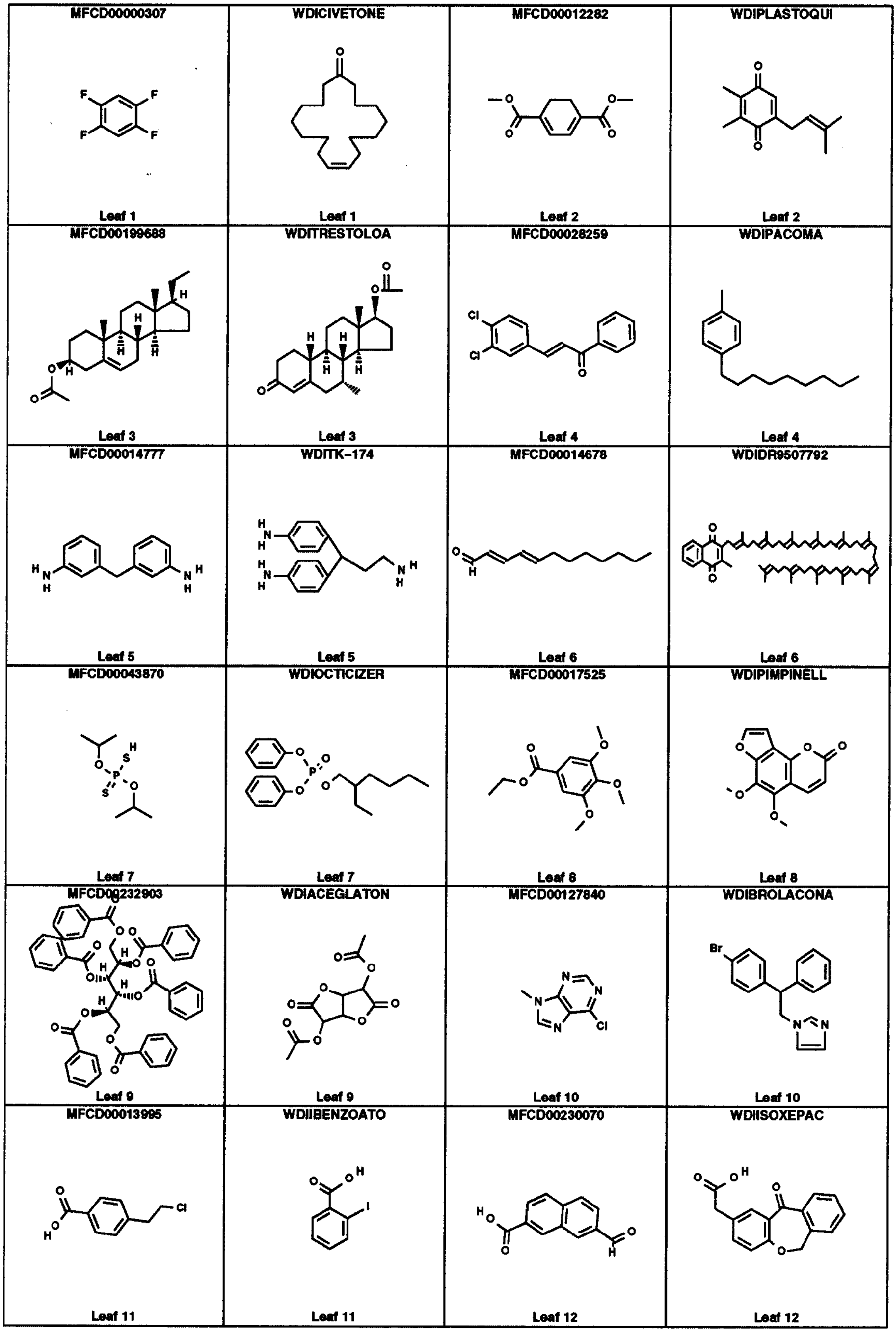
Minimizing the Prediction Error for Drugs. So far only the overall error rate on an independent test set has been used as a measure to compare the performance of different decision trees: It was assumed that the lower the error rate, the better is the corresponding model. A more detailed picture can be obtained by analyzing the different types of errors that have been made: how many drugs have been classified as nondrugs, and vice versa. These two types of error correspond to false negative and false positive classifications of drugs, respectively. A convenient way to display this information is a so-called confusion matrix: On the diagonal of the confusion matrix, the percentage of correctly classified drugs and nondrugs is given, respectively. The two off-diagonal entries are showing the misclassification rates, i.e. the percentage of drugs classified as nondrugs, and vice versa. In Table 2, the confusion matrix for the tree with 266 leaf nodes and a test set error of 20.9% is shown. When comparing the different error rates, it can be seen that they are not the same for the two classes: 18.9% of the nondrug compounds are predicted to be potential drugs, whereas 23.0% of the known drugs are classified as nondrugs. This situation is actually contrary to the one desired for compound acquisition or the design of combinatorial libraries: For these purposes the number of missed potential drugs (false negatives) should be as small as possible, even at the expense of an increased error in the classification of nondrugs.

This problem can be solved by weighting the two possible types of errors differently during the generation of a decision tree, i.e. by assigning different “costs” to false negatives and false positives. If costs of 5 and 1 are assigned to false negatives and false positives, respectively, the classification of a drug as a nondrug will be 5 times as “expensive” during the construction of a decision tree as the classification of a nondrug as a drug. In other words, the number of misclassified drugs (i.e. the number of false negatives) will be reduced at the cost of an increased number of misclassified nondrugs (i.e. the number of false positives).

In order to study the influence of different misclassification costs on the error rates, trees with misclassification costs between 1 and 10 for false negatives and fixed costs of 1 for false positives were generated using two different sets of training parameters (see Figure 6). The training parameters of $M = 2$, $CF = 8$ and of $M = 56$, $CF = 1$ correspond to the two trees previously discussed in section *Selecting the Most Predictive Model*. Accordingly, the error rates for misclassification costs of 1 are identical to the results obtained there without explicit misclassification costs.

With increasing costs for the incorrect classification of drugs the rate of false negatives is decreasing as expected. However, this is only possible due to an increased misclassification rate of nondrugs. In the extreme cases with misclassifications costs greater than 7, the rate of incorrectly classified drugs can approach 0, but only because *all* compounds are classified as drugs. Correspondingly, the misclassification rate of nondrugs approaches 100% in these cases. On the basis of the results of Figure 6, we chose a misclassification rate of 3 as a good trade-off between loosing too many true drugs and accepting too many nondrugs: With this setting only 10% of the drugs are erroneously classified as nondrugs, but it is still possible to correctly identify about 60% of the nondrugs.

The influence of the training parameters on the prediction error was studied again, but this time with a cost of 3 associated with the misclassification of a drug as a nondrug (cf. section *Selecting the Most Predictive Mode*). In this situation the quality of the model cannot be assessed by


MFCD00000307
Leaf 1
WDICIVETONE
Leaf 1
MFCD00012282
Leaf 2
WDIPLASTOQUI
Leaf 2
MFCD00199688
Leaf 3
WDITRESTOLOA
Leaf 3
MFCD00028259
Leaf 4
WDIPACOMA
Leaf 4
MFCD00014777
Leaf 5
WDITK-174
Leaf 5
MFCD00014678
Leaf 6
WDIDR9507792
Leaf 6
MFCD00043870
Leaf 7
WDIOCTICIZER
Leaf 7
MFCD00017525
Leaf 8
WDIPIMPINELL
Leaf 8
MFCD00232903
Leaf 9
WDIACEGLATON
Leaf 9
MFCD00127840
Leaf 10
WDIBROLACONA
Leaf 10
MFCD00013995
Leaf 11
WDIIBENZOATO
Leaf 11
MFCD00230070
Leaf 12
WDIISOXEPAC
Leaf 12

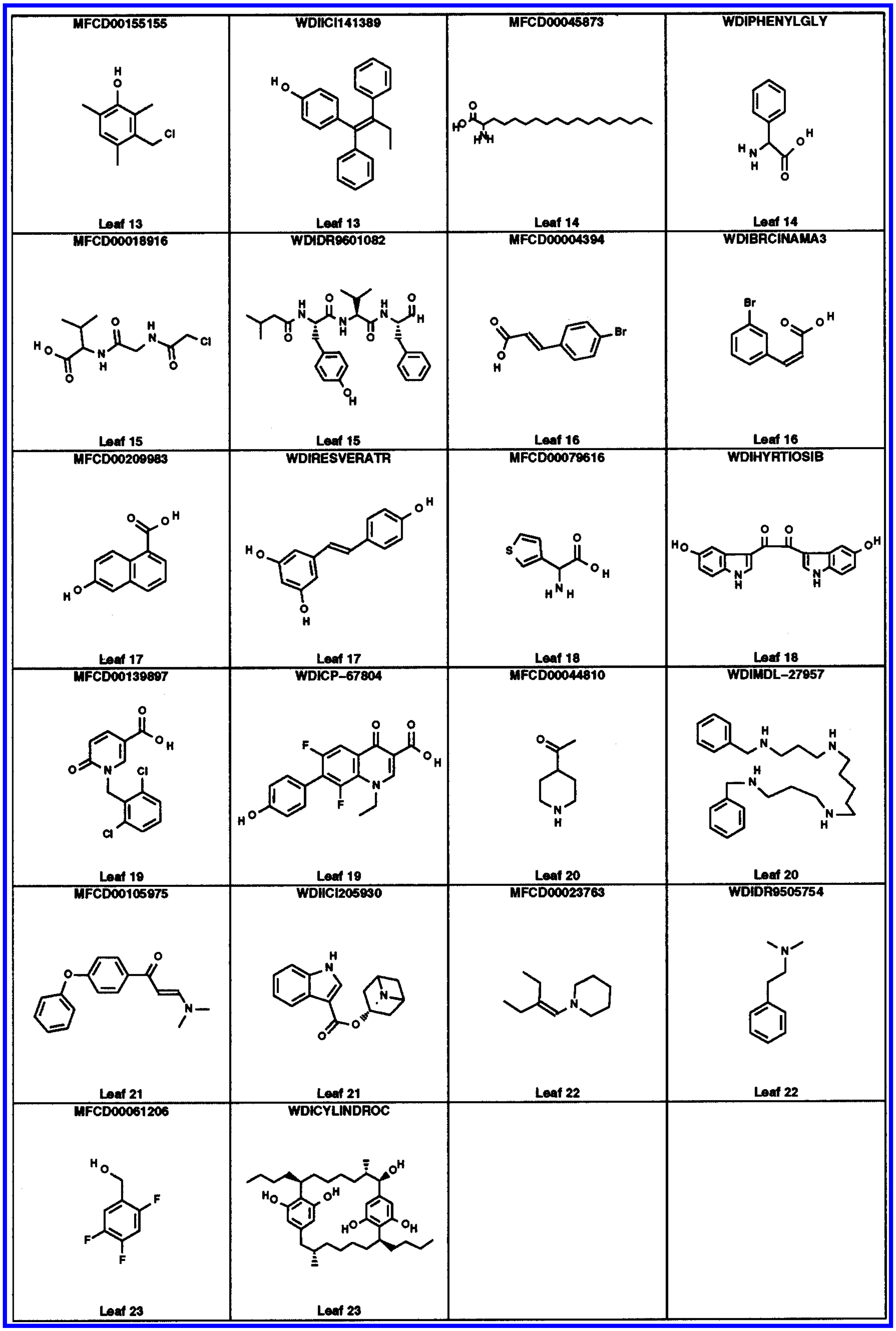


Figure 5. Representative structures from the ACD and WDI that are mapped onto the different leaves of the decision tree shown in Figure 4. The compound name and the number of the corresponding leaf node are given above and below each structure, respectively. The compound name starts with MFCD for compounds from the ACD and with WDI for compounds from the WDI.

looking at the prediction error alone, but a decision has to be made on the rate of false negatives one is willing to tolerate. When this number has been decided upon, then the model with the lowest rate of false positives for the chosen rate of false negatives can be selected. Figure 7 shows the relationship between the two misclassification rates for 2340

Table 2. Confusion Matrix of the Predictions Made by the Decision Tree with 266 Leaf Nodes

	predicted %	
	as D	as N
true class D	77.0	23.0
true class N	18.9	81.1

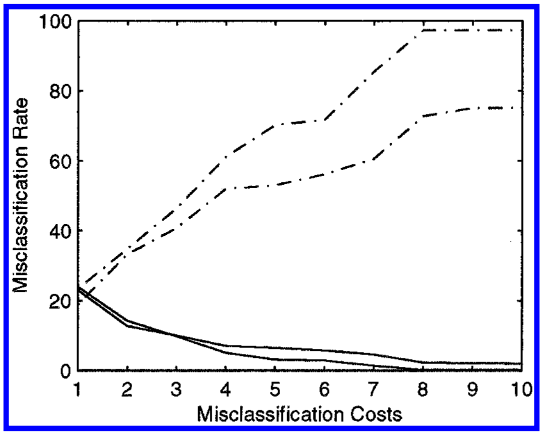


Figure 6. Misclassification rate as a function of misclassification costs. The solid lines show the rate of false negatives; the dotted lines, the rate of false positives. The two sets of curves correspond to two different sets of training parameters for the decision tree.

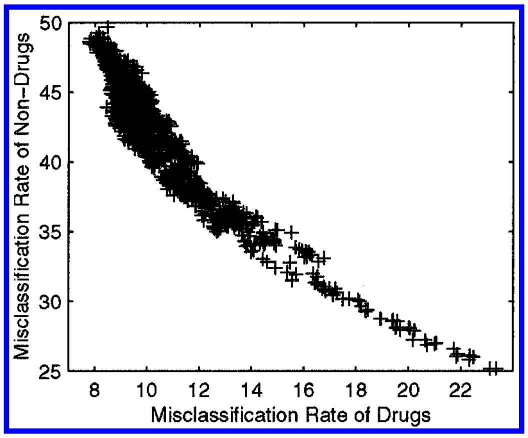


Figure 7. Relationship between the misclassification rate of drugs (rate of false negatives) and the misclassification rate of nondrugs (rate of false positives) calculated with misclassification costs of 3 for the incorrect classification of drugs.

different settings of the training parameters *M* and *CF*.

Figure 7 reveals that both the rate of false negatives and false positives vary to quite some extent with the training parameters of the decision tree. All the points in the figure represent specific trade-offs between the misclassification rate of drugs and nondrugs, respectively, but only those points with a minimum rate of false positives for a given rate of false negatives need to be considered further. From among these models a tree with a misclassification rate of 9.0% for drugs and 42.5% for nondrugs, respectively, was chosen for further investigations. The tree has 142 leaf nodes and was trained with the following parameters: $CF = 16$, $M = 13$. The prediction error of this tree estimated by a third, independent validation data set is 26.0%; the misclassification rates of drugs and nondrugs on the validation data set are 8.9 and 43.1%, respectively.

In order to analyze the differences between the models with and without misclassification costs, a tree ($CF = 1$, $M = 34$) with the same number of leaf nodes as the one shown in Figure 4 was identified among the 2340 trees that have been trained with misclassification costs of 3. A comparison of the two trees, the one with (Figure 8) and the one without misclassification costs (Figure 4), can show how similar the two models are and whether the same descriptors are important in both of them. The test set error of the tree with misclassification costs is 29.1% in comparison to 23.9% for the tree without misclassification costs. Although the overall error is much higher for the former tree, its rate of false negatives is only 8.5% in comparison to 23.9% false negatives for the latter tree.

The error rate based on the training set is again color-coded in the tree trained with misclassification costs of 3 shown in Figure 8. In comparison to the one shown in Figure 4, quite some increase of the error rate can be noticed when following the main branch of the tree. Furthermore, some leaf nodes classified as drugs show a high error rate, e.g. the first leaf node in the upper left corner (leaf 1) contains only 459 drugs, but 643 nondrugs, corresponding to an error rate of 58.3%. Still, as a consequence of the specified misclassification costs, this and comparable nodes with more nondrugs than drugs are classified as drugs. This increases the error rate, but the misclassification costs at these nodes are still decreased.

It is interesting to note that not all of the leaf nodes in Figure 8 show such a high error rate: The seven leaf nodes labeled as nondrugs have an average error rate of only 13.6%. This is due to the fact that the specified misclassification costs favor very much the classification as a drug, so that all borderline cases will be classified as drugs. Consequently, the nodes labeled as nondrugs in this figure supply a much better description of the typical features of nondrugs than the leaf nodes of the tree from Figure 4. Especially leaf nodes 8 and 15 in Figure 8 with more than 1000 compounds each show the importance of aromatic carbon atoms for the classification of nondrugs: Compounds in leaf 8 need to have more than six aromatic carbon atoms with one hydrogen attached, compounds in leaf 15 more than two of that type of carbon atoms.

For the identification of drugs the same descriptors are important as in the tree previously trained without misclassification costs: Almost three quarters of all the drug compounds in the training data set are recognized in Figure 8 by simply testing for the presence of hydroxyl, tertiary amino, secondary amino, carboxyl, phenol, or enol groups (leaves 21, 20, 19, and 18 in Figure 8). As a consequence of the aforementioned classification of borderline cases as drugs, this general structure is even more obvious in Figure 8 than it is in Figure 4.

Combining Several Decision Trees. The analysis of the different decision trees described in the previous sections shows that the prediction error can change to quite some extent between different leaf nodes of the same tree. This is an indication that there might be special subclasses in the training data set that are much harder to predict than others—a situation where boosting can be beneficial. Therefore, two boosted trees were generated, one without and one with penalizing the misclassification of drugs. In both cases, the parameters identified as optimum for the nonboosted trees were used, i.e., $M = 2$, $CF = 8$ for the boosted tree without misclassifications costs and $CF = 16$, $M = 13$ for the boosted

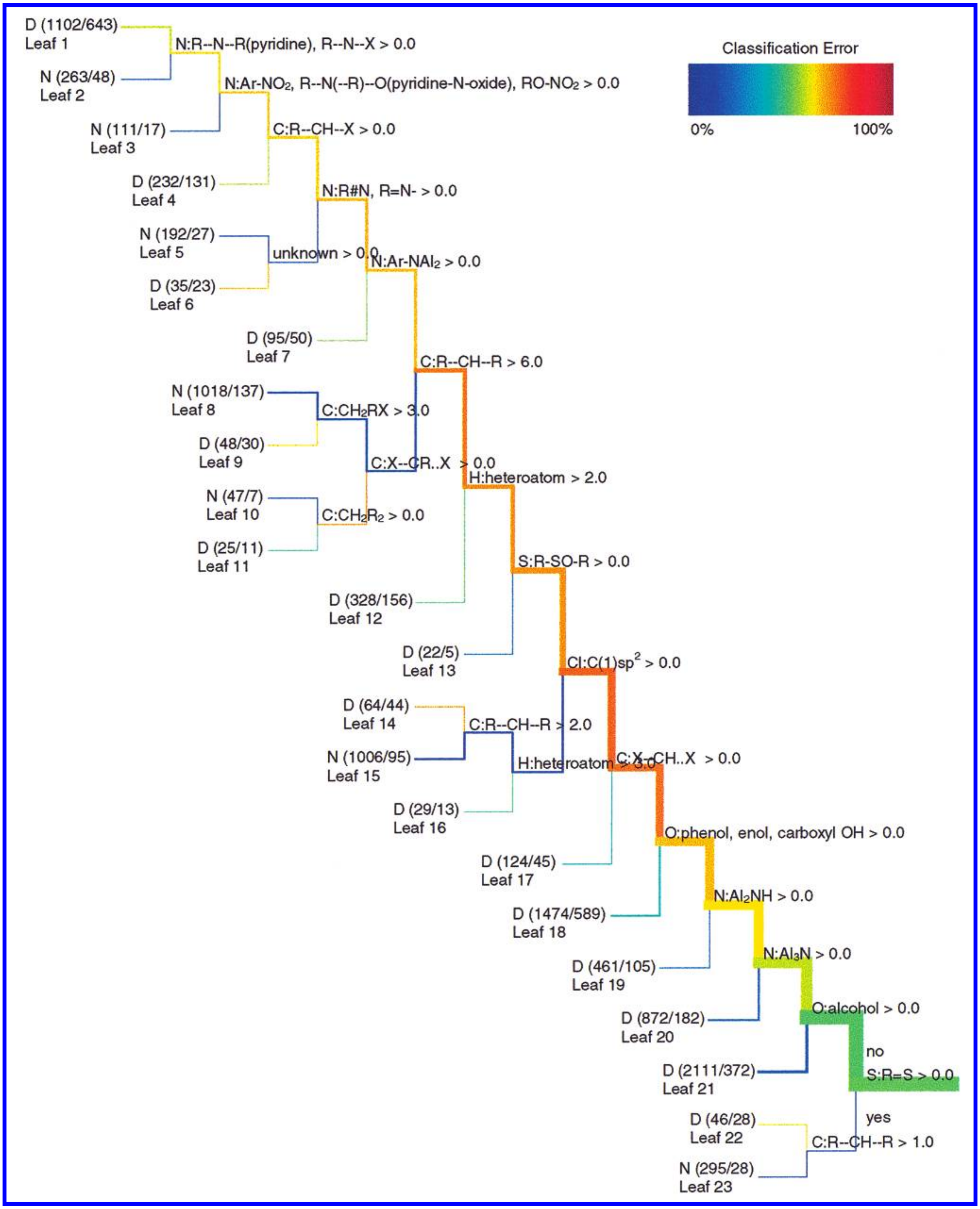


Figure 8. Simplified decision tree to predict whether a compound is likely to act as a drug or not. The tree has been trained with misclassification costs of 3 for incorrectly classifying a drug as a nondrug. Each leaf of the tree is labeled either with D for drugs or with N for nondrugs. Behind the label the number of compounds in that leaf and the number of classification errors are given. The line width indicates the number of compounds; the classification error of the training data set is color-coded (blue = 0%, green = 50%, and red = 100% error). The descriptors and tests used for the classification are given above each nonterminal node. The following abbreviations are used: R represents any group linked through a carbon; X represents any heteroatom; Al and Ar represent aliphatic and aromatic groups, respectively; −, =, and # represent single, double, and triple bonds, respectively; - - represents aromatic bonds as in benzene or delocalized bonds such as in nitro groups; •• represents aromatic bonds such as the CN bond in pyrrole; C(*n*)sp^x represents a carbon with formal oxidation number *n* and hybridization sp^x.

Table 3. Comparison of the Predictions Made by the Normal and the Boosted Decision Tree

	normal tree		boosted tree	
	test set	validation set	test set	validation set
prediction error (%), no misclassification costs	20.9	21.4	18.0	17.4
misclassified drugs (%), misclassification costs of 3	9.0	8.9	8.7	8.1
misclassified nondrugs (%), misclassification costs of 3	42.5	43.1	34.7	34.3

Table 4. Predictions Made for Drugs, a Database of Known Drugs from the Literature

	prediction, %			
compds classified as drugs in	normal tree	tree with misclassification costs	boosted tree	boosted tree with misclassification costs
complete Drugs database	77.2	90.2	81.7	91.9
subset of Drugs database not in the WDI	83.1	92.8	85.8	94.2

tree with misclassifications costs of 3, and the number of boosted trees was set to 10 (*NB* = 10). Table 3 gives an overview over the results.

In all cases a substantial improvement could be achieved using boosted trees. The prediction error was reduced by ca. 3% for trees without misclassification costs, i.e. a relative improvement of ca. 15% could be obtained. In the case of trees with penalized misclassification of drugs, the number of false negatives was only reduced moderately, but at the same time the misclassification rate of nondrugs has decreased from ca. 42% to ca. 34%.

These results show that boosting is an effective way to improve the prediction accuracy with and without penalizing the misclassification of drugs. One drawback though is that the boosted trees cannot so conveniently be visualized and interpreted anymore: Each model now consists of 10 different trees and furthermore, during voting each of the 10 trees has different weights for different predictions.

Application to Databases. In order to test the generated models, the four decision trees described in the previous sections (with/without penalizing misclassification of drugs and boosted/nonboosted) have been applied to three corporate structural databases at Organon: *Drugs*, an inhouse collection of known drugs extracted from the literature, *Chembase*, a database of historical Organon compounds, and *Diverse*, a diverse collection of compounds from one of Organon's screening programs.

The Drugs database contains more than 12 000 known drugs extracted from literature, paying special attention to the therapeutic areas Organon is interested in. Since the Drugs database has been generated independently from the WDI, it represents another test case for the predictive power of the generated decision trees. The overlap of 52% between the WDI and the Drugs database is relatively low, probably due to the above-mentioned bias in the Drugs database. Table 4 shows the percentage of compounds classified as potential drugs by the four different models.

These classification rates are in total accordance with the results obtained with the test and validation sets in the previous sections: About 23% of the drugs are misclassified as nondrugs by the normal decision tree. This misclassification rate can be decreased to about 18% using a boosted tree. If costs of 3 are specified for the misclassification of drugs as nondrugs, both the boosted and the nonboosted trees show a misclassification rate lower than 10%. Interestingly, the subset of compounds that can only be found in the Drugs database, but not in the WDI, is predicted to be between 2 and 6% more druglike than the complete Drugs database. This is certainly a consequence of the structural bias in the Drugs database (vide supra); i.e. certain structural classes that are difficult to predict are underrepresented among the compounds that are exclusive to the Drugs database.

Table 5. Predictions Made for Chembase, a Database of Historical Organon Compounds

	prediction, %			
compds classified as drugs in	normal tree	tree with misclassification costs	boosted tree	boosted tree with misclassification costs
complete Chembase	63.2	85.0	64.4	81.8
subset of end products in Chembase	77.7	90.3	78.7	90.3
subset of precursors in Chembase	55.2	82.0	56.4	77.1

Chembase is a database of compounds synthesized by medicinal chemists at Organon. It contains both compounds that were especially designed for specific therapeutic targets and synthetic precursors for these compounds. The former compounds should in general be classified as potential drugs by the decision trees—the structures were especially designed to be drugs. The remaining structures, most of them synthetic precursors, should be the less druglike the farther away they are synthetically from the intended end product of the synthesis. Therefore, the percentage of druglike compounds among the precursors should be on average lower than among the end products. Table 5 gives the percentage of compounds classified as drugs in Chembase and in the two subsets, end products and precursors.

As expected, over 60% of the compounds in Chembase are classified as drugs; more than 80% if misclassification of drugs as nondrugs is penalized. The results for precursors and end products show notable differences. Using the normal decision tree, about 78% of the end products are predicted as drugs, whereas only about 55% of the precursors are classified as being druglike. The difference of 23% between the two parts of Chembase shrinks to about 8% if misclassification of drugs is penalized. This is reasonable, as borderline cases will be classified as drugs if the misclassification of drugs is penalized, and therefore a comparatively higher percentage of the precursors will be classified as drugs. Interestingly, the figures for the end products in the above table closely resemble the results obtained for the entire Drugs database, indicating again the druglike nature of the end products in Chembase.

The third database that has been studied consists of a collection of compounds that were screened for several different biological activities. Since the compounds in the database have been selected in a way that the diversity of the collection is maximized, a significant part of it will not be druglike. Biological activity is the crucial requirement for a compound to become a drug. Therefore, the percentage of druglike compounds among those that showed activity in

Table 6. Predictions Made for Diverse, a Diverse Collection of Compounds Screened with Several Assays

	prediction, %			
compds classified as drugs in	normal tree	tree with misclassification costs	boosted tree	boosted tree with misclassification costs
complete Diverse database	41.5	62.7	40.1	59.1
subset with activity in at least one assay	60.0	80.5	60.5	78.0
subset with no activity in any assay	33.0	54.5	30.7	50.4

at least one assay should be higher than among those that showed no activity at all. The results for these database are summarized in Table 6.

As expected, only 40 or 60% of the compounds in the Diverse database are considered to be potential drugs— depending on whether misclassification of drugs has been penalized or not. However, the rate of potential drugs is about 30% higher among those compounds that showed activity in at least one assay than among those that showed no activity at all. If the predictions of the *boosted tree with misclassification costs* were used for selecting compounds from the Diverse database, only 60% of the compounds would have been selected for screening. In doing so, 50% of the compounds that showed no activity at all could have been avoided, but also 22% of those compounds that showed some activity would have been omitted from screening.

DISCUSSION

We have introduced a model that can discriminate between potential drugs and nondrugs. It is able to correctly classify 82.6% of a validation data set that has been selected independently from the training data set. The predictions of the model can be used to guide the purchase or selection of compounds for biological screening. The design of combinatorial libraries can be supported by comparing virtual libraries generated with different sets of reagents in order to optimize the overall druglike character of the final library.

Regarding the intended application, it is desirable to keep the number of incorrectly classified drugs, i.e. false negatives, as low as possible, even if the number of false positives increases in consequence. Therefore, a second model was developed that penalizes the misclassification of nondrugs. The resulting model correctly classifies 91.9% of the drugs in the validation set, but unavoidably the rate of false positives increases to 34.3% at the same time.

The model has been trained using 10 000 compounds from the ACD and the WDI, with 5000 compounds randomly selected from each of the two databases. This selection of the training data set is based on the assumption that the two databases mostly consist of drugs and nondrugs, respectively. By removing those structures from the ACD that could be found in both databases, the ACD and the WDI, some measures were taken to reduce the number of incorrect classifications in the data sets used. However, there are still nondrugs to be found in the processed WDI (e.g. pharmaceutic aids, reagents, etc.) and druglike compounds in the processed ACD. This unavoidable overlap between the two databases defines a lower limit for the prediction error that can be achieved.

Still, the more fundamental question, whether the generated model can really distinguish between potential drugs and nondrugs, or only between compounds from the WDI and from the ACD, cannot be answered by looking at the above prediction errors. Therefore, we applied the two models to three different in house compound collections: Drugs, a set of known drugs from the literature that is supposed to be somewhat similar to the WDI, Chembase, a set of historical Organon compounds, and Diverse, a set of compounds that has been screened with several assays. In each of the three cases, the compounds known to be drugs or druglike were classified significantly more often as potential drugs than the remainder of the compounds. Consequently, the models were able to generalize the information contained in the training set and are not especially fitted to the contents of the WDI and ACD.

The models have been developed using decision trees as implemented in the program package C5.0. In contrast to other classification methods, the resulting decision trees are not only of high predictive power but can also be interpreted by examining the decision trees. An analysis of the simplified decision trees representing the models with and without penalizing false negatives revealed important features that distinguish druglike structures from nondrugs: Just by testing the presence of hydroxyl, tertiary or secondary amino, carboxyl, phenol, or enol groups, already three quarters of all drug compounds can be recognized. The nondrugs, on the other hand, are characterized by their aromatic nature with a low content of functional groups besides halogens. Although these conclusions were drawn on the basis of simplified decision trees, they are nevertheless quite valid, as the simplified trees have the same general structure as the trees with the best prediction results and a predictive power that is reduced by only ca. 5%.

This ease of interpretation is also what distinguishes our approach from other recently published methods to differentiate between drugs and nondrugs that are based on neural networks.[9,10] Our as well as these methods achieve a prediction error of approximately 20%.[22] However, both authors state that they could not analyze the trained neural networks in order to better understand the characteristics that distinguish drugs from nondrugs. On the basis of the analysis of the decision trees, we were able to identify a few, surprisingly simple features that explain the most significant differences between drugs and nondrugs. Although these distinctive features seem to be trivial (presence/absence of functional groups that may form hydrogen bonds), they only become so once they have been identified.

ACKNOWLEDGMENT

We thank J. Sadowski for interesting discussions on the subject of this paper.

Supporting Information Available: Table giving a complete listing of all 121 descriptors used (2 pages). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Gallop, M. A.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gordon, E. M. Application of Combinatorial Technologies to Drug Discovery. 1. Background and Peptide Combinatorial Libraries. *J. Med. Chem.*

1994, *37*, 1233−1251. Gordon, E. M.; Barrett, R. W.; Dower, W. J.; Fodor, S. P. A.; Gallop, M. A. Application of Combinatorial Technologies to Drug Discovery. 2. Combinatorial Organic Synthesis, Library Screening Strategies, and Future Directions. *J. Med. Chem.* **1994**, *37*, 1385−1401.

(2) The entire March 1996 issue of *Acc. Chem. Res.* is dedicated to combinatorial chemistry.

(3) Thompson, L. A.; Ellman, J. A. Synthesis and Applications of Small Molecule Libraries. *Chem. Rev.* **1996**, *96*, 555−600.

(4) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring Diversity: Experimental Design of Combinatorial Libraries for Drug Discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.

(5) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. W.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.

(6) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(7) Menard, P. R.; Lewis, R. A.; Mason, J. S. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(9) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.

(10) Ajay; Walters, W. P.; Murcko, M. A. Can We Learn To Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(11) Gillet, V.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(12) In ref 10, Ajay et al. also mention the use of decision trees to distinguish between drugs and nondrugs. Since, in their setup, the results with decision trees are always worse than with neural networks, they did not pursue this approach.

(13) Quinlan, J. R. C4.5: *Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.

(14) *World Drug Index*, version 2/96; Derwent Information: London, UK, 1996.

(15) Available Chemicals Directory, version 2/96; MDL Information Services: San Leandro, CA, 1996.

(16) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three-Dimensional Structure Directed Quantitative Structure−Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(17) Hawkins, D. M.; Young, S. S.; Rusinko, A., III Analysis of a Large Structure-Activity Data Set Using Recursive Partitioning. *Quant. Struct.-Act. Relat.* **1997**, *16*, 296−302.

(18) Quinlan, J. R. Bagging, boosting, and C4.5. In *Proceedings of the 13th American Association for Artificial Intelligence National Conference on Artificial Intelligence.* AAAI Press: Menlo Park, CA, 1996; pp 725−730.

(19) *C5.0*, release 1.08a; RuleQuest Research Pty Ltd., St Ives NSW, Australia (http://www.rulequest.com/; e-mail, quinlan@rulequest.com).

(20) C5.0 somewhat heuristically assumes that the number of training set errors in a node is binominally distributed. Then, the upper confidence limit of the binominal distribution for a given confidence level *CF* is used as the prediction error rate at that node.[13]

(21) Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141−142.

(22) The cross-validated prediction error reported by Ajay et al.[10] amounts to only ca. 10%. On a data set derived from the MDDR, another database of known drugs that is independent from the training set used, they achieved a prediction error between 16 and 23%.

CI990266T