A Scoring Scheme for Discriminating between Drugs and Nondrugs

Jens Sadowski* and Hugo Kubinyi

Combinatorial Chemistry and Molecular Modelling, ZHF/G - A 30, BASF AG, D-67056 Ludwigshafen, Germany

Received October 6, 1997

A scoring scheme for the rapid and automatic classification of molecules into drugs and nondrugs was developed. The method is a valuable new tool that can aid in the selection and prioritization of compounds from large compound collections for purchase or biological testing and that can replace a considerable amount of laborious manual work by a more unbiased approach. It is based on the extraction of knowledge from large databases of drugs and nondrugs. The method was set up by using atom type descriptors for encoding the molecular structures and by training a feedforward neural network for classifying the molecules. It was parametrized and validated by using large databases of drugs and nondrugs (169 331 molecules from the Available Chemicals Directory, ACD, and 38 416 molecules from the World Drug Index, WDI). The method revealed features in the molecular descriptors that either qualify or disqualify a molecule for being a drug and classified 83% of the ACD and 77% of the WDI adequately.

Introduction

With the advent of combinatorial chemistry and high throughput screening as methods for the generation and testing of large numbers of molecules in drug design, also a number of computer-based methods focused on these working areas appeared recently (for an overview see ref 1). Besides specific book-keeping problems with combinatorial libraries or screening results in database software, they are mostly centered on similarity and diversity as criteria for compound selection either for choosing building blocks for combinatorial libraries or for purchasing compounds from external sources for screening purposes.

Three questions are frequently asked: (1) Which subset of a combinatorial library, of a set of building blocks, or of a molecular database spans the most diverse chemical space? (2) Which subset of a given compound library fills most effectively the "holes" in an existing in-house database? (3) Which subset is most similar to a given lead compound and gives therefore the highest chance for finding new hits? To answer these questions, a number of molecular descriptors as well as a number of statistical methods have been investigated. In some cases it could be shown that a compound selection guided by similarity criteria could significantly enlarge the portion of biologically active compounds compared to a random selection.

However, a much simpler question has not yet been considered: Which subset of a compound library is most "drug-like" and gives the highest chance for finding new screening hits? This question arises when choosing molecules from rather heterogeneous sources as, e.g., from the catalog of a supplier of chemicals. There are obvious criteria for excluding compounds that are not suited as, e.g., reactive chemicals.[2] These can easily be implemented in an automatic filter program that excludes compounds containing certain substructures as, e.g., acid halogenides or isocyanates. Furthermore, a medicinal chemist applies a huge amount of intuition to classify compounds into potential drugs and nondrugs. He/she implicitly considers possible drug−receptor interaction sites as well as more general criteria such as bioavailability, toxicity, and mutagenicity. He/she often applies rather complex expert rules that cannot be easily coded into a simple substructure filter. Therefore, this second approach for selection is most often manually performed by medicinal chemists—a rather tedious and time-consuming task. Moreover, each individual chemist prefers or disfavors certain chemical classes, he/she is familiar with, accordingly the selection will be biased. Thus, a general and objective filter that automatically distinguishes between compounds showing a certain potential for being a drug and other, not suited compounds, is highly desirable. The literature on computational approaches for this task or related topics is rather sparse.[3] Therefore, a new scoring scheme for discriminating between drugs and nondrugs was developed. It is based on the assumption that typical drugs have something in common that other compounds lack. The method was parametrized and tested by using large available structural databases containing some hundreds of thousands of compounds.

Materials and Methods

General Outline. A scoring scheme was established for the classification of molecules into drug-like and nondrug-like compounds. The basic idea is to obtain a set of general and objective rules about structural features that are obviously essential for drugs. This knowledge must be implicitly contained in large collections of either basic chemicals or drugs. It can be assumed that a collection of known drugs can reveal structural features qualifying a compound for being a drug. On the other hand, a collection of nondrugs will give some hints for features disqualifying a compound for being a drug. Two large databases were chosen for the parametrization of the method. One database contained drugs and drug candidates and the other one was a heterogeneous mixture of currently available chemicals. A neural network was trained to predict to which database a given compound belongs, i.e., whether it could be a drug or not.

Preparation of the Data. The Available Chemicals Directory (ACD)[4] and the World Drug Index (WDI)[5] were chosen as databases for the parametrization of the method. The ACD is a collection of compounds that are often enough not drug-like, e.g., intermediates and reactive compounds, whereas the

S0022-2623(97)00677-8 CCC: $15.00
Published on Web 08/01/1998

Table 1. Characteristics of the Databases

	ACD	WDI
compounds	240 347	50 472
valid	217 963	50 471
suited[a]	187 570	44 839
duplicates[b]	12 426	6 423
shared[c]	5 813	5 813
final[d]	169 331	38 416

[a] After applying the substructure filter. [b] After removing counterions and solvents and after neutralizing formal charges. [c] After removing duplicates. [d] After removing duplicates from both databases and after removing shared compounds from the ACD.

WDI mainly contains a large and diverse collection of known drugs. The molecular structures were filtered and normalized in order to fulfill the following criteria:

1. The records are valid, i.e., they contain the connection table fields and there are no obvious errors in the structure description.
2. The compounds pass a rejection filter that removes chemically reactive or otherwise not suited compounds.[6]
3. Counterions and solvent molecules are removed in order to obtain single-compound records.
4. Charges at acidic and basic groups are neutralized by adding or removing protons.
5. Duplicates within each individual database are removed.
6. Compounds shared by both databases are removed from the ACD.

The databases are characterized by the figures given in Table 1. Criterion 3 was applied in order to adequately handle compounds differing only in additional fragments. Criterion 4 was applied in order to overcome structural differences caused by different protonation states. After these corrections, both databases contain a considerable amount of duplicates, i.e., compounds either multiply present in the original databases or differing only in counterions, solvents, or the protonation state. There is also a rather large overlap of compounds in both databases. This gives rise to the assumption that the ACD contains already a considerable amount of drugs or drug-like molecules. One has to be aware about this bias in the data. Anyway, to avoid as much false classification as possible, all compounds contained in both databases were classified as drugs and consequently removed from the ACD. In addition, one could try to remove further drug-like molecules from the ACD either manually or by finding close analogues of the WDI compounds applying a standard similarity search technique. We refrained from such approaches since this would introduce another bias into the data. The manual search for drug-like molecules would be biased by the favors of the chemist performing it. The similarity search would be based on the assumption that the similarity definition of the particular approach is correlated with biological activity. This cannot be simply proven. We all know examples where the replacement of a certain substituent of a drug produces a still very similar but completely inactive analogue.

The preparation of the data was accomplished by assigning score values for the "drug-likeness" of 0 and 1 to the ACD and WDI compounds, respectively. For the calibration of a system for predicting these scores, subsets of 5000 compounds were randomly extracted from both databases giving a training set of 10 000 compounds in total.

Molecular Descriptors. Ghose and Crippen have established a system of atom types for encoding organic molecules.[7] The counts of these 120 atom types in a molecule have been used as molecular descriptors in the present study. Although tailored for a rather different purpose—the prediction of the lipophilicity parameter log *P*—these atom types give a rather detailed description of organic molecules. We used only a subset of 92 atom types populated at least 20 times in the training set of 10 000 compounds. This descriptor set is something like an extended molecular formula and gives a total of 92 input values describing one molecule.

A number of alternative descriptors as global as log *P* or molecular weight or as detailed as Daylight fingerprints were also investigated. They were found to be less well suited as the Ghose/Crippen atom types (data not shown). A possible explanation is that we need a compromise between too global descriptors and too detailed descriptors in order to encode enough information but maintain sufficient freedom for generalization.

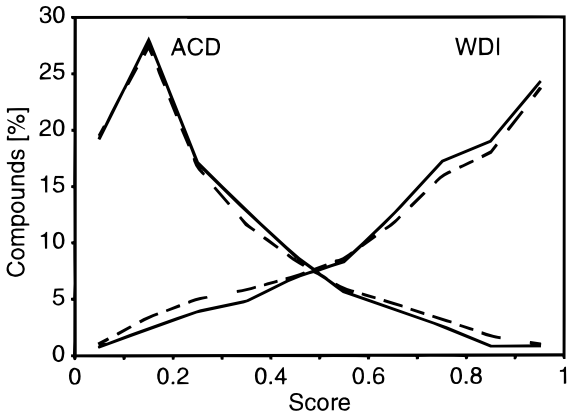

Figure 1. Distribution of predicted scores for the training sets (solid lines) and for the complete databases (dashed lines).

Neural Network Training. The SNNS program[8] was used for all neural network operations. Feedforward nets were constructed that consist of 92 input neurons (Ghose/Crippen atom types), five hidden neurons, and one output neuron (score). All layers were totally connected, resulting in a total of 465 weights. The net was trained with the molecular descriptors as input values and the scores as output values. For technical reasons, all input and output values were scaled between 0.1 and 0.9. The net was trained following the "backpropagation with momentum scheme" as implemented in SNNS. The training was performed over 2000 cycles with a learning rate of 0.2 and a momentum term of 0.1. The training dataset was shuffled before each cycle, i.e., the training data were in each training cycle presented to the neural network in a new order. Test runs showed that the training process achieved sufficient convergence with these parameters.

Results and Discussion

Reproduction and Prediction of the Data. A 92 × 5 × 1 feedforward network was trained with a training dataset of 5000 ACD compounds and 5000 WDI compounds. The trained net was used to predict the scores of the compounds in the training set and in the whole ACD and WDI databases. Figure 1 shows the distribution of the predicted scores for these datasets. Solid lines are assigned to the training sets of 5000 ACD compounds and 5000 WDI compounds. Dashed lines represent the prediction of the total of 169 331 ACD compounds and 38 416 WDI compounds from this model. There is a clear discrimination between the drugs in the WDI and the heterogeneous compounds of the ACD.

Separating drugs and nondrugs according to a borderline set at a scoring value of 0.5, 83% of the ACD compounds and 77% of the WDI compounds were classified correctly. Moreover, the net trained by only 10 000 randomly selected compounds used in the training set (solid lines) succeeded to classify the much larger number of about 210 000 compounds of the complete databases (dashed lines) with nearly the same quality. Thus, the training sets were representative selections from the whole databases, and the good classification of the training sets did not result from a mere over-

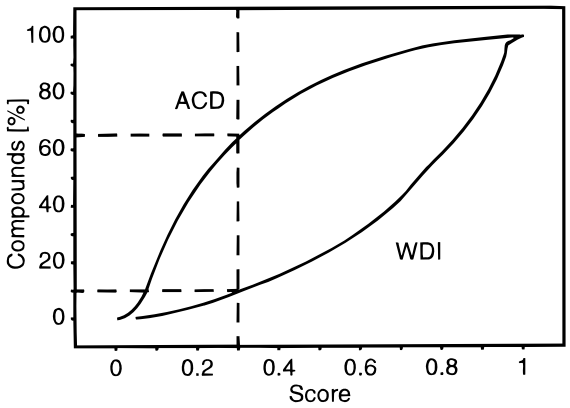

Figure 2. Sorted score distributions for the complete databases. Dotted lines indicate a threshold score of 0.3 and the excluded database portions resulting from this score.

training of the neural net. On the other hand, there remains a considerable number of drugs in the WDI with rather small scores. This sheds some light on the limitations of the method. The approach suffers mostly from the misclassification of a considerable amount of compounds in the ACD training set. Only those ACD compounds with exact matches onto the WDI (5813 shared compounds) could be removed. Thus, a significant number of undetected drugs or drug-like compounds are still contained in our reduced ACD. These have consequently been assigned an incorrect score value of 0.1. This fact will bias the training of the network. Since there is no way of assigning automatically the correct score values to all of the training compounds one has to accept a threshold of false positives as well as false negatives. Despite this limiting aspect of the present approach, the discriminative power of the trained network is rather astonishing.

It would be interesting to trace back why the trained network is capable of classifying some molecules as drugs and others as nondrugs. Unfortunately, a complex nonlinear model as represented by a neural network cannot be interpreted straightforward in order to extract some rules from it. We are not aware of any method suitable for analyzing trained neural networks in such ways. A work-around could be to feed small fragments or functional groups into the trained neural network in order find those that contribute most to the score. But the approach is based on whole drug molecules on one side and basic chemicals that in principle could serve as building blocks for the drugs on the other side. In addition, the neural network is a nonlinear model and one cannot expect mere additivity of the fragment scores. Accordingly, simple tests did not reveal any hint on favorable or unfavorable fragments or functional groups (data not shown). The discrimination is certainly based on a more complex network of conditions for a drug-like molecule such as "two groups A, one group B, and no group C" or "one group A and one group C", etc.

Misclassified Compounds. To give an impression of compounds that were not classified correctly, 10 arbitrary pairs of misclassified molecules from the ACD and the WDI are presented. Figure 3 shows five ACD compounds with a score greater than 0.7 along with very similar compounds of the WDI. Figure 4 shows five arbitrary compounds from the WDI with a score of less

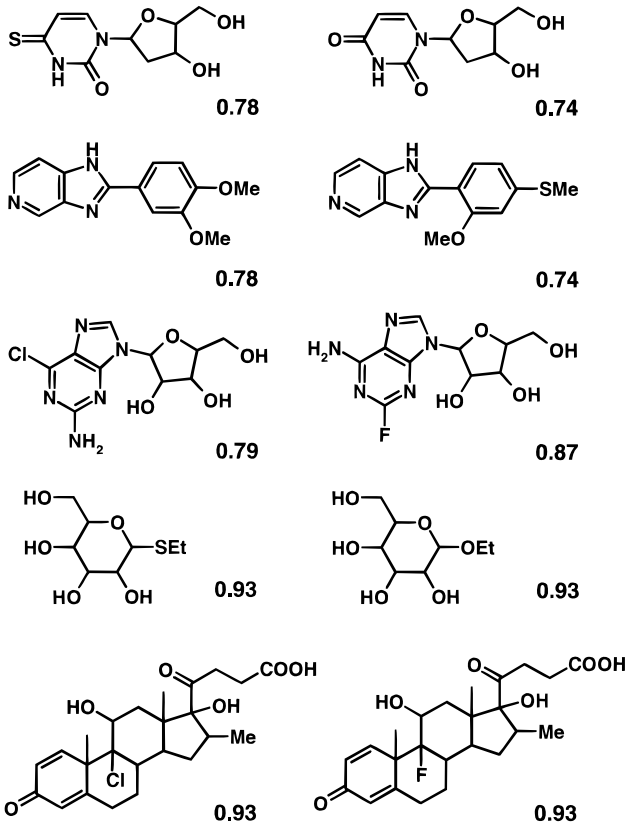

Figure 3. Examples of ACD compounds having scoring values greater than 0.7 (left-hand side) along with very similar WDI compounds (right-hand side).

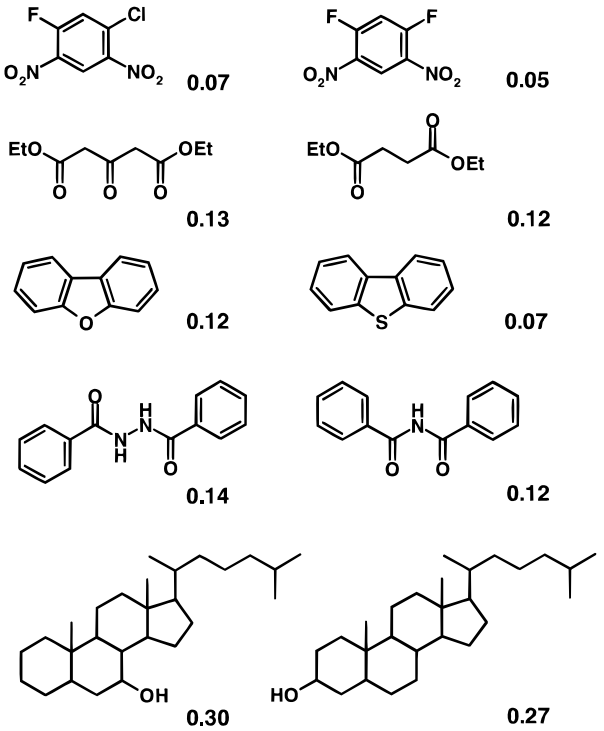

Figure 4. Examples of ACD compounds (left-hand side) along with very similar WDI compounds (right-hand side) having scoring values less than 0.3.

than 0.3 along with very similar ACD compounds. This gives an idea where the limitations of such an approach are. All pairs of compounds differ only in marginal structural details, and the neural net could not dis-

Table 2. Calculated Score Values for a Number of Best-Selling Drugs[9]

name	score	name	score
ranitidine	0.78	lovastatin	0.89
enalapril	0.82	cimetidine	0.72
fluoxetine	0.53	omeprazole	0.85
aciclovir	0.64	cefaclor	0.67
simvastatin	0.80	ceftriaxone	0.97
co-amoxiclav		estrogenes	
amoxicillin	0.80	estrone	0.62
clavulanic acid	0.68	equilin	0.73
diclofenac	0.40	cyclosporin	0.84
ciprofloxacin	0.93	beclometasone	0.82
nifedipine	0.76	famotidine	0.65
captopril	0.82	salbutamol	0.93
diltiazem	0.80	sertraline	0.65

criminate between them (we will not discuss here which of the ACD compounds might possibly be drugs).

Selecting Compounds According to the Scoring Value. Figure 2 illustrates some implications of the scoring scheme for compound selection. The accumulated scoring values of both databases are shown. One can arbitrarily define threshold values for the score and select all compounds from a database with a score beyond a particular threshold. If we decide to tolerate a loss of 10% of the drugs in the WDI (lower horizontal line), this would require a threshold value of about 0.3 (vertical line), at the same time excluding 67% of the ACD (upper horizontal line). The application of such rejection criteria allows one to restrict purchase and screening of new compounds to the most promising ones. Accordingly, this means an optimization of the efforts for finding new lead structures. The definition of the threshold between drugs and nondrugs remains with the user. It could be guided by the portion of drugs from the WDI database that would be lost when applying a distinct threshold score.

It must be emphasized that the system presented here gives only rather rough predictions whether compounds are drug-like or not. As can be seen from the score distribution of the World Drug Index, there is also a considerable amount of WDI compounds assigned to rather low scoring values in our approach. Thus, the scoring scheme should not be used for the evaluation of single compounds. It should only be applied to exclude a portion of molecules with a very low score from purchase or testing in order to enhance the portion of potential drugs for screening purposes. Beyond a certain threshold value, all remaining compounds should be ranked equally and further selections should be performed either randomly or diversity-driven. One cannot expect any correlation between the score values and eventual biological activities in certain biological tests.

Score Values of Top-Selling Drugs. To validate the method and to give an impression for the meaning of the score values, the predicted scores for a number of best-selling drugs are given (Table 2).[9] They all are assigned to scoring values greater than 0.5 with one exception (diclofenac, 0.40), and all of them are beyond

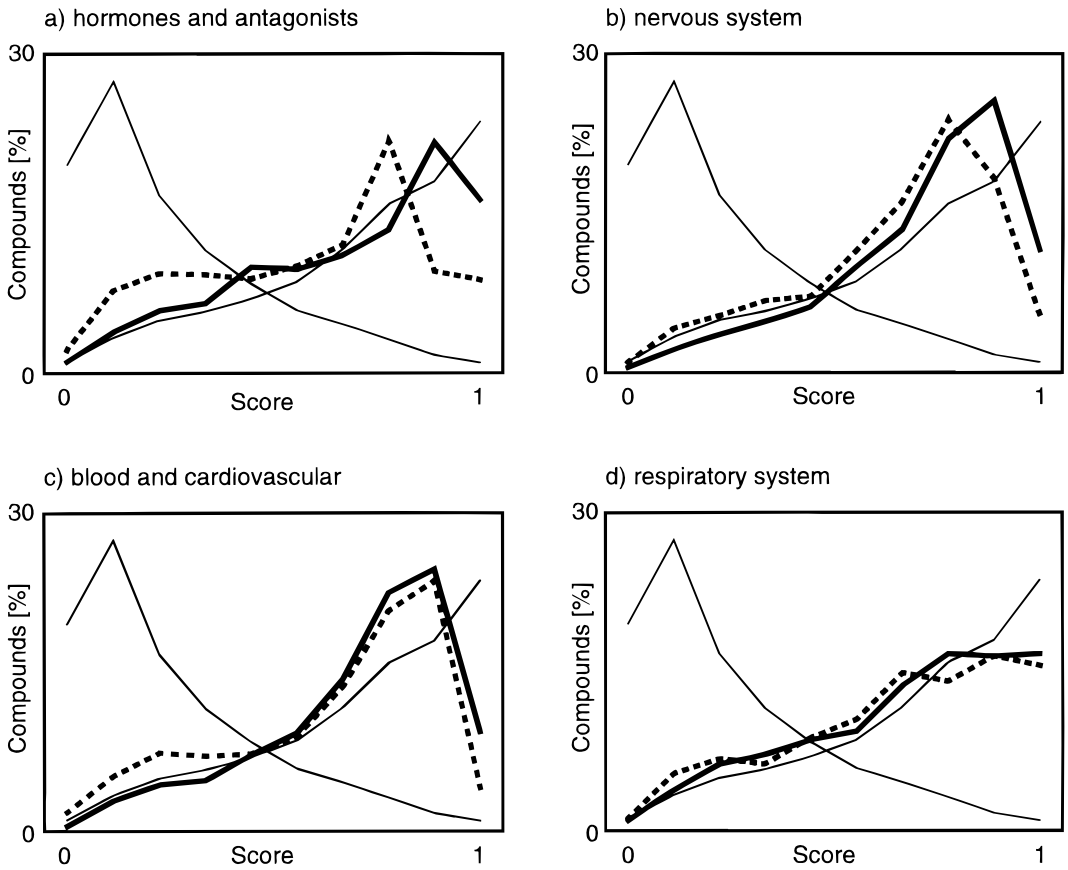

Figure 5. Score distributions for different indication areas (see text) as obtained from different training sets. For each indication area, one training set included the compounds of a certain indication area (solid lines), whereas the other one did not include them (dotted lines). For comparison, also the distributions for the whole ACD and WDI databases are given (thin lines, cf. Figure 1).

Table 3. Prediction of Scores for Molecules of Various Indication Areas with Models Including and Excluding These Compounds[a]

		score > 0.5 (%)		score > 0.3 (%)	
indication area	compounds	incl	excl	incl	excl
hormones and antagonists	3077 (8%)	73	62	89	79
nervous system	3844 (10%)	83	75	94	89
blood and cardiovascular	4347 (11%)	80	71	92	86
respiratory system	1023 (3%)	73	72	89	87

[a] For each indication area are given the following: the number of compounds in the total database, the percentages of compounds with a score greater than 0.5, and the percentages of compounds with a score greater than 0.3.

the threshold value of 0.3. Thus, the scoring scheme classified all of them correctly as drugs.

Robustness of the Method. In an additional study, the robustness and predictive power of the method was investigated. A serious objection against the present approach might be that it was trained by using the currently available pharmaceutical knowledge. Thus, it could fail in correctly handling compounds in new, not yet considered indication areas or chemical classes. Such future areas are of course not represented by the World Drug Index or anywhere else. The question is whether the approach is able to generalize on a level beyond of certain drug types or chemical classes and if it can handle new areas correctly or not. We simulated this situation by successively excluding four complete indication areas from the training set. These were in particular hormones and antagonists (8% of the WDI), drugs acting on the nervous system (10%), drugs acting on the blood and cardiovascular system (11%), and drugs acting on the respiratory system (3%). Figure 5 shows plots of the score distributions obtained by neural nets trained with and without these compound sets. Due to the exclusion, the curves are slightly shifted toward lower scores. But most compounds in the four different indication areas are more or less correctly classified as drugs, independently of whether a particular indication area has been included into the training set or not.

Table 3 lists for these subsets the number of compounds in the total database, the percentages of compounds with scores greater than 0.5 and 0.3 obtained by models including and excluding the corresponding compounds, respectively. The comparison of these values before and after the exclusion shows no significant increase in the percentage of wrongly classified compounds. For a threshold of for example 0.3, the increase varies between 2% (respiratory system) and 8% (hormones and antagonists). Thus, it appears that drug molecules have some general structural features in common that differentiate them from the vast majority of organic molecules. The approach taken in this study appears capable to extract these features, even if some compound classes active in particular indication areas were systematically excluded from the model derivation process.

Timings. The method for scoring molecular databases is rather fast. For example, processing the ACD and WDI databases takes 32 and 10 min, respectively, on an R10000 175 MHz Silicon Graphics CPU. This means, the scoring of one single molecule takes about 0.01 s on average—a rather insignificant amount of computer time.

Conclusions

A fast automatic scoring scheme was established and parametrized for the discrimination between drugs and nondrugs. It succeeded to classify correctly 83% of the contents of the Available Chemicals Directory as nondrugs and 77% of the World Drug Index as drugs. The method can be used for selecting compounds for purchase or biological testing. It could be shown that the approach revealed certain features in molecules that either qualify or disqualify them to be a drug. Besides that, the robustness and predictive power of the approach was demonstrated. It was shown that the method is able to correctly predict drugs of whole indication areas that were excluded from the training set.

It is possible to apply the same approach to similar problems. We found comparable results for the discrimination between crop protection compounds and basic chemicals based on in-house data and the ACD (data not shown). Other applications could be in toxicity or mutagenicity prediction.

Acknowledgment. We gratefully acknowledge inspiring discussions with our colleagues U. Abraham, T. Mietzner, G. Paul, J. Delzer, and P. Eckard. We thank G. Klebe (University of Marburg) and an anonymous referee for many hints and remarks on the manuscript.

Note: After finishing this work we were notified of a similar work by Ajay and co-workers.[10] They used ISIS-keys and Bayesian neural networks for distinguishing between drugs and nondrugs in the ACD and the CMC databases with similar good results.

References

(1) Warr, W. A. Combinatorial Chemistry and Molecular Diversity. An Overview. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 134−140.
(2) Rishton, G. M. Reactive compounds and *in vitro* false positives in HTS. *Drug Discovery Today* **1997**, *2*, 382−384.
(3) Gillet, V. J.; Willett, P.; Bradshaw, J. Development of Bioactivity Profiles for Use in Compound Selection. 211th ACS National Meeting, March 24−28, 1996, CINF 066.
(4) ACD: Available Chemicals Directory; Version 2/96, MDL Information Systems, 1996.
(5) WDI: World Drug Index; Version 2/96, Derwent Information, 1996.
(6) The following rules for removing unsuited compounds were applied: Molecular weight less than 150 or greater than 1000. Metal-containing compounds (Li, Na, K, Mg, Ca, Zn, and Al are allowed as cations). More than six halogen atoms or multivalent halogens (except in counterions). Hydrocarbons. Reactive acid derivatives: halogenides, anhydrides, cyanides. Peroxides, isocyanates, azides, diazonium compounds, isonitriles. Phosphanes. Silicates.
(7) 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.
(8) SNNS: Stuttgart Neural Network Simulator; Version 4.0, University of Stuttgart, 1995.
(9) SCRIP No. 2040, July 7, 1995; p 23.
(10) Ajay, Vertex Pharmaceuticals, personal communication.

JM9706776