Gradual in Silico Filtering for Druglike Substances

Nadine Schneider, Christine Jäckels, Claudia Andres, and Michael C. Hutter*

Center for Bioinformatics, Saarland University, Building C 7.1, P.O. Box 15 11 50, D-66041 Saarbruecken, Germany

Received September 19, 2007

The suitability of decision trees in comparison to support vector machines for the classification of chemical compounds into drugs and nondrugs was investigated. To account for the requirements upon screening virtual compound libraries, schemes for successive filtering steps with gradual increasing computational cost are outlined. The obtained prediction accuracy was similar between decision trees and support vector machine approaches for the applied compound data sets. By using rapidly computable variables such as druglikeness indices, XlogP, and the molar refractivity, at least 39% of the nondrugs can be filtered out, while retaining more than 83% of the actual drugs. Computationally more demanding descriptors such as specific substructure queries and quantum chemically derived variables can be postponed to subsequent classification schemes for the reduced set of compounds, whereby up to 92% of the nondrugs can be sorted out without loosing considerably more drugs. Using all available computed descriptors simultaneously in the first step did not yield significantly better results. Furthermore, the generated decision trees are used to derive guidelines for the design of druglike substances. The numerical margins found at the branching points suggest several criteria that separate drugs from nondrugs: a molecular weight higher than 230, a molar refractivity higher than 40, and the presence of one or more rings as well as one or more functional groups. Also reported are additionally required parameters to compute values for XlogP, SlogP, and the molar refractivity of boron and silicon containing compounds.

INTRODUCTION

Computer-aided drug design has become an indispensable tool in the pursuit of innovative and new pharmaceutical drugs. Prediction methods and algorithms are not limited to the individual components of pharmacokinetics alone, namely absorption, distribution, metabolism, elimination, and toxicity.[1,2] The prediction accuracy of such individual ADME properties is, however, dependent on the availability of sufficient experimental data. In the course of screening virtual compound libraries it is therefore desirable to sort out compounds efficiently that are per se unsuitable, before more specific prediction models are applied. Moreover, the challenge of selecting potentially suitable compounds for in vitro screening from the vast chemical space has led to approaches that predict the druglikeness by the means of a single index.[3−6] The results of these studies have, however, shown that an unequivocal assignment is rather difficult, even using neural nets to generate similar indices.[3,7−10] Since separation of drugs from nondrugs can be regarded ultimately as a classification problem, other machine learning techniques have also been applied and compared to each other.[7,11−17] However, even support vector machines that are known to be among the most accurate methods for classification tasks have so far not provided fully satisfactory results.[17−20] This can be, at least in part, explained by the implicit difficulties that arise from the data set.

At first glance, the assignment of a chemical compound as drug or nondrug may seem trivial. One would consider those substances that are marketed as drugs and that have been previously assigned to a distinct pharmaceutical category as being drugs. Consequently, all compounds that have not being assigned to a category should be nondrugs. Browsing through substance collections, i.e., the Merck index,[21] however, shows that numerous compounds exhibit potential pharmaceutical functions such as antibacterial activity, although they are not assigned to a corresponding category for various reasons including toxicological issues. Are such compounds drugs, nondrugs, or something in between that can be referred to as druglike? Conversely, typical antineoplastic drugs possess severe mutagenic properties, e.g., DNA-alkylating agents. Naively, one would not consider such compounds as typical drugs. If we, however, exclude this category due to the toxicological implication, we would also have to exclude a whole series of other categories as well, for example acetylcholinesterase inhibitors that are functionally related to insecticides. In turn Gálvez et al. have raised the question how to assign natural products that exhibit a nonpharmaceutical function such as vitamins.[12] At this point it should have transpired that the intended classification scheme is subject to some inherent difficulties, and thus the results must be analyzed considering these aspects. As a consequence we will have to expect a larger number of falsely classified compounds in the mentioned and presumably further problematic categories.

Although the final algorithmic solution of the classification problem itself may be available at some stage, the major question in drug design remains: What structural features cause a substance to be a suitable drug or at least a druglike compound. Answers to this questions should guide the

* Corresponding author phone: (49)-681-302-64178; fax: (49)-681-302-64180; e-mail: michael.hutter@bioinformatik.uni-saarland.de.

10.1021/ci700351y CCC: $40.75
Published on Web 02/13/2008

synthetic design of new potential drugs. Analyses of databases containing drugs have provided a series of answers from different points of views. Molecular skeletons and side chains that most frequently appear in drugs have been collected by Bemis and Murcko.[22,23] Substructure analysis is frequently used to detect potentially suitable compounds.[24,25] Muegge et al. found that a certain number of functional groups is mandatory, whereas nondrugs are typically underfunctionalized.[26] Gedeck and Willet have furthermore reviewed analysis of structure−activity relationships in high-throughput screening data especially under the aspect of classification of compounds into drugs and nondrugs.[27] The statistical distribution of certain features between drugs and nondrugs has also been exploited to derive druglikeness indices: Ertl used organic substituents and fragments, whereas Hutter focused on atom pair combinations.[5,6] Limits for orally administered drugs regarding molecular weight and the number of hydrogen-bond donors and hydrogen-bond acceptors as well as lipophilicity were suggested by Lipinski and co-workers.[28] To account for 70% of all drugs, Oprea suggested refined ranges for these descriptors and included additional criteria, i.e., the number of rings and rotatable bonds.[29] A much more extensive set of rules was proposed by Xu and Stevenson.[4] Considering also nonorally administered drugs, Ghose et al. yielded margins that comprised the preferred 50% and a less stringent range of 80% of druglike substances taking similar variables into account.[30]

The latter methods correspond to a gradual filtering of the considered substances based on a series of molecular descriptors. A similar step-by-step separation is also performed by decision tree algorithms. In this context Wagener and van Geerestein found that three-quarters of all drugs can be correctly assigned based on the occurrence of six chemical groups.[11] The outcome of the mentioned filtering approaches is, however, 2-fold. The results can readily be used for the synthesis of similar compounds because the initial selection of the available decision criteria was based on the knowledge of medicinal chemists. Rare or newly introduced fragments of chemical structures may therefore be classified false or missed. Likewise, the majority of per se unsuitable compounds can be ruled out by the presence of certain chemical groups that render a substance being either reactive, toxic, or difficult to synthesize. Hann et al. as well as Flower have presented similar substructures in the form of SMARTS strings whereby such compounds can be detected.[31,32] Considering specific toxicological effects, Wang et al. have collected a series of fragments from the RTECS (Registry of Toxic Effects of Chemical Substances) database that contains similar information.[33] These substructures are, however, rather complex and furthermore found in many safe drugs as well. Their use in the context of separating potentially druglike compounds from ordinary chemicals is thus limited. Moreover, the practical application of substructure queries is computationally more demanding than other simple descriptors and thus less feasible for the screening of a vast number of substances, e.g., complete virtual libraries.

Even more time-consuming is the use of quantum chemically derived descriptors, e.g., those related to the molecular electrostatic potential, although their application is promising and a suitable extension to conventional variables.[14] To circumvent this bottleneck upon in silico screening for druglike compounds, it is necessary to narrow the number of compounds to a feasible size. This can be achieved by setting up a gradual filtering strategy whereby the whole set of compounds is filtered according to rules based on descriptors that can be computed rapidly from the molecular structure, as suggested previously.[6] For those molecules that are classified as being druglike by this first filter, the necessary number of substructure queries are performed that constitute the second set of rules. The number of compounds that pass this second filtering step will now be reasonably small enough to allow quantum chemical calculations to be carried out to derive the third and last set of rules.

In principle it is possible to use any of the mentioned classification algorithms for this purpose. Two of them, namely support vectors machines and decision trees, are, however, particularly useful in this context and will be applied in this study for the following reasons: Support vectors machines produce very accurate classifications but require an appropriate selection of descriptors as input similar to neural nets. Conversely, decision trees perform an appropriate choice from even a very large set of input descriptors due to their recursive partitioning strategy. Thus their classification scheme can be interpreted easily, and the most significant descriptors usually appear at early branching points.[11,34,35] Furthermore, rule based selection criteria can be derived straightforwardly from decision trees and be converted into guidelines for the design of druglike compounds. In turn, it is of interest to compare the results of such recursive rule based classification schemes to those obtained from support vector machines that imply a simultaneous fit of all data. The available descriptors for this purpose comprise variables that can be generated rapidly, e.g., druglikeness indices and simple selection rules, substructure fragments expressed by SMARTS strings that require more computing time, and eventually quantum chemically derived variables that involve a suitable 3-dimensional molecular structure.

It is apparent that any classification method will yield more or less false predictions for drugs and nondrugs. In the context of gradual filtering it is of course more favorable to retain as many drugs as possible in each step and to carry on more nondrugs, rather than to lose a considerable number of potential drugs right from the beginning. So far, the mentioned studies have throughout reported better prediction accuracies for drugs than for nondrugs, regardless of the applied classification method. Thus we wittingly exploit this circumstance to achieve a gradual enrichment of drugs.

METHOD

1. Compound Data Sets. The chemical substances used in this study are an extended set of the data used previously.[6] Compounds were collected from the Merck Index,[21] G. Milne's compilation of drugs,[36] and an older commercial vendor catalog of chemicals.[37] Analysis of the compound properties (see Results) showed a very similar distribution compared to larger data collections such as the MACCS-I Drug Data Report (MDDR), the Comprehensive Medicinal Chemistry (CMC), and conversely the Available Chemical Directory (ACD) for nondrugs. Only compounds containing the elements C, N, O, F, Cl, Br, I, S, P, Si, and B were

considered. Deprotonated acids (e.g., carboxylates, sulfonates, and phosphates) were converted into their neutral form by replacing Na^+ and K^+ by hydrogen to ensure a net neutral charge. Likewise, quaternary nitrogens were "neutralized" by adding Cl^-. Inner salts were kept unchanged. Thus a net neutral charge of the compounds is ensured. This is a necessity for the performed quantum chemical calculations in order to avoid a number of problems. First, the dipole moment is not independent of its origin for compounds with a net charge unequal to zero. Second, the computed molecular electrostatic potential becomes imbalanced for compounds with a charge excess, thus causing distinct outliers for these molecules. Furthermore, mixtures of two or more compounds were discarded to ensure that only single molecules are present. Likewise, HCl and water were removed from the entries where necessary. This procedure is in line with previous studies.[3,11,26]

Classification of the compounds into drugs and nondrugs was performed under consideration of the following criteria:

1. Nondrugs must not be assigned to a pharmaceutical category but can be a diagnostic aid, pharmaceutical aid, dye, pigment, vitamin, solvent, surfactant, insect repellent, sun screen, ultraviolet screen, sweetener, or flavoring. In contrast to other studies where reactive and toxic compounds are removed from the data set, we included such substances to capture their characteristics, e.g., those of acaricides and herbicides.[3]

2. Drugs are characterized by affiliation to a typical pharmaceutical category but excluding inhalative administered compounds. Except antineoplastics and antivirals those compounds that primarily exhibit a toxicological or mutagenic function were omitted, e.g., pesticides, insecticides, rodenticides, miticides, and plant growth inhibitors.

Applying these selection criteria yielded a total of 3117 pharmaceutical drugs and 2238 nondrugs. Further 51 pharmaceutical compounds that were featured in the journal "Drugs of the Future" between January 2005 and December 2006 were taken as an external validation set. These druglike compounds are about to enter clinical trials and are given along with their classification results in Table 7. Similar SMILES strings of all compounds are provided as Supporting Information. The training and test sets of drugs and nondrugs will also be part of the upcoming CoEPrA (Comparative Evaluation of Prediction Algorithms) modeling competition, to allow further comparison regarding other partitioning algorithms.[38] The compounds were further partitioned into nonoverlapping training and test sets whereby the respective test sets comprised 10% of all compounds. Partitioning was performed with respect to the chemical diversity. For this purpose a series of fingerprintlike descriptors was computed indicating the presence of certain substructures, e.g., methoxy groups, sulfonamides, esters, azides, heterocyclic 5-membered rings, and fused ring systems occurring in common drugs.[22] These binary descriptors were generated on the basis of similar SMARTS strings applied to the SMILES strings of the respective compounds using the "obgreb" command of Open Babel.[39] These SMARTS strings (a total of 168) are given in Table S1 of the Supporting Information. Furthermore values for XlogP,[40] SlogP,[41] and the molar refractivity[41] were computed using an in-house Python program. Additional parameters required for boron and silicon containing compounds developed in our laboratory are given in the Supporting Information. Splitting of the compounds into a training and a test set was performed on the basis of a cluster analysis using an in-house Python program. Clusters were generated by applying single linkage using the City-Block (Manhattan) distance.[42] The 168 SMARTS strings as well as the molecular weight, XlogP, and the molar refractivity were considered as descriptor space. Compounds being closest to one of the centroids of the respective clusters were chosen for the test sets and therefore should be representative samples. Principal component analysis of the descriptors showed that the first component explained more than 91% of the total variance, except for the test set of nondrugs (84%). The sum of the first two components explained more than 99% of the total variance for all compound sets.

Further computed descriptors that were also available to the classification algorithms, but not used for clustering, comprised the number of hydrogen-bond donors and hydrogen-bond acceptors, the total number of halogen atoms, the number of carboxylic acids as well as other acids, the respective count of hydroxy, amino, NO_2, sulfoxy, sulfonyl, SO_3, nitrile, CF_3, CCl_3, and ester groups as well as the total sum of these functional groups, the number of unsuitable groups according to Flower[32] also including occurrences of matching substructures according to the SMARTS strings compiled from Hann et al.,[31] Oprea,[29] Rishton,[43] and Anzali et al.[13] (see Table S2 in the Supporting Information), the respective number of 3-, 4-, 5-, and 6-membered rings as well as aromatic 5- and 6-membered rings, the total number of rings, the number of violations of Lipinski's rule,[28] the 50% and 80% criteria of drugs by Ghose, Viswanadhan, and Wendoloski,[30] Oprea's druglike criteria at 70%,[29] and Hutter's druglikeliness index.[6] Quantum chemical descriptors were obtained from semiempirical AM1 calculations using a modified version of the program package VAMP.[44,45] Compounds were energetically optimized to a gradient norm below 0.1 kcal mol^{-1} $Å^{-1}$ using the default Eigenvector Following algorithm.[46] For particularly large compounds with a molecular weight of 600 or above, the convergence criterion was loosened to 0.2 kcal mol^{-1} $Å^{-1}$. Descriptors were generated from the respective output files using an in-house PERL script to obtain appropriate input data for the decision tree algorithm. The full list of all descriptors and SMARTS strings used is provided as Supporting Information.

Compounds for further test and validation sets of drugs and nondrugs were taken from Murcia-Soler et al.[9] Not included were substances labeled as nondrugs for those pharmaceutical indications reported in the Merck Index, namely bergenin, bostrycoidin, collinomycin, cusparine, enviroxime, fumigatin, hadacidin, honokiol, magnolol, metitepine, nybomicin, phaseolin, thidiazuron, viridicatin, vitamin K_5, cinnabarine, fuscin, noformicin, sparassol, tiafur, usinic acid, and vitamin A_2.[21]

2. Decision Tree. The algorithmic concept of the applied decision tree that was developed in our lab has been described in full detail in an earlier publication.[34] The greedy separation strategy is based on recursive partitioning and creates an iterative branching topology in which the branch taken at each intersection is determined by a rule related to a single descriptor of the molecule. Finally, each terminating leaf of the tree is assigned to a class. To avoid excessive partitioning, here the maximum branching depth was limited

to 5 since further branching did not show significant improvement (see Results). Likewise, this avoids "pruning" of the decision trees to remove branching points that separate only a few compounds. In cases where two or more descriptors yield an identical separation at a given branching point, the separation is based on that variable that shows the smallest variance among its range of data. Our previous results have shown that this strategy yields a higher prediction accuracy for test sets, since it accounts for the expected clustering of data points in different classes.[34,35] In contrast to regression equations where descriptors showing large variances are of importance to derive a continuous quantity, decision trees (and likewise support vector machines) target a maximum separation of classes and thus depend on descriptors that account for the clustering of data. The variance of a variable does, however, not express clustering effects among its data range sufficiently: The same numerical variance might reflect either a broad distribution around an average value or two clearly separated peaks. Thus a conventional principal component analysis to detect suitable variables is not helpful in this context.

3. Support Vector Machine. The support vector machine as implemented in the e1017 package[47] of R[48] was applied for performing classifications. A linear kernel function with default settings was chosen to avoid overfitting that might arise from nonlinear functions (see Discussion).

4. Timing. Generation of the first and largest decision tree A (4819 compounds, 249 descriptors) took about 2.5 h on an Intel Pentium IV processor running at 2.6 GHz. Smaller decisions trees for the same number of compounds but comprising a reduced number of descriptors took 25 and 4 min, respectively, on the same machine. Predictions based on the rules derived by these decision trees are performed within less than 10 s for the full set of data. The support vector machines took 1.2 h applying all 249 descriptors and 11 and 2 min, respectively, for smaller numbers of variables.

RESULTS AND DISCUSSION

For the intended gradual filtering the following strategy was pursued: First, the optimal choice of descriptors was determined that allows a step-by-step filtering with increasing computational cost. Thus, rapidly computable variables are applied for the whole set of compounds, while descriptors that require an increased computational effort only have to be generated for those substances that pass preceding filtering steps. Each filtering step is carried out by a specific decision tree using the similar choice of descriptors (see Figure 1). For comparison, the filtering steps have also been carried out using support vector machines. Furthermore, the numerical margins obtained for the descriptors being used by the decision trees were compared to existing guidelines regarding the design of druglike substances.

Several sets of descriptors were computed for the compounds of the data set that are used as input for the two classification methods, namely decision trees and linear support vector machines. These descriptor sets are denoted A to F, and the similar classification schemes based on them are named likewise (see Figure 1). While set A comprises all 249 available descriptors (see Tables S1 and S3 of the Supporting Information), the other sets B to F contain only a subset of the descriptors. The first decision tree (A) was

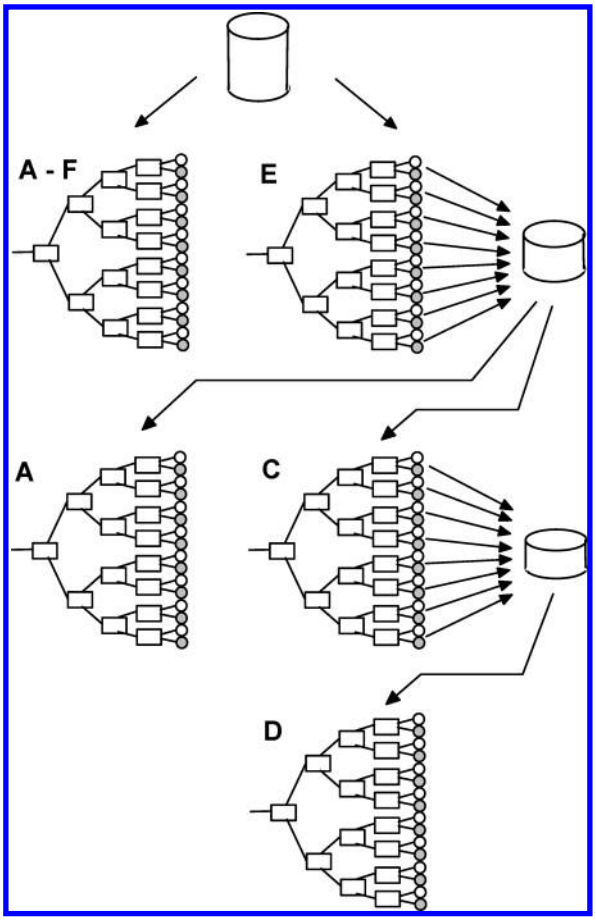

Figure 1. Workflow of the gradual filtering. To perform the splitting of the initial data set into drugs and nondrugs either decision trees or support vector machines are used. They are denoted A to F depending on the underlying descriptor set, respectively (cf. Table 1).

generated for the 4819 compounds in the training set using the full set of computed variables to reveal the relevant descriptors that separate drugs from nondrugs. Descriptors used at branching points are given in Table 1 and explained briefly in Table 2. The very first descriptor used by the algorithm is surprisingly the molecular volume (box A in Figure 2). By applying this first separator already more than 77% of all compounds can be correctly assigned as being drugs or nondrugs. Analysis of the intercorrelation between the available descriptors showed, however, strong pairwise correlation between the molecular volume and the molecular weight ($r = 0.955$) and the molar refractivity ($r = 0.983$). This can easily be explained since molecular weight, molecular volume, and likewise molar refractivity increase almost linearly with the number of atoms in a molecule. Anyhow, the molecular volume is chosen as the first separator since it allows a better classification according to the layout of the decision tree strategy that chooses that descriptor that allows the best possible separation at the respective branching point. Thus compounds with a molecular volume of more than 191 Å^3 are more likely to be drugs than nondrugs. Interestingly, the correlated variables (molar refractivity and molecular weight) appear at later branching points (C2 and D4) with separating margins of 40 and 139, respectively. This value for the molecular refractivity is identical to the lower margin for druglike substances derived by Ghose et al. and is also found in other decision trees (see

Table 1. Criteria Used at Branching Points in the Decision Trees[g]

criteria	tree A[a]	tree B[b]	tree C[c]	tree D[d]	tree E[e]	tree F[f]
A	MOLVOL	MW	NRING	MOLVOL	MW	MW
B1	M092	M092	M092	PSANEG	HUTTER	XlogP
B2	FUNGR	HUTTER	NROTB	VMINUS	HUTTER	FUNCGR
C1	M084	M093	M093	G1GEO	SO2	NROTB
C2	MR	MR	NROTB	COHESI	MR	NRING
C3	PSA	M063	M161	ESPMIN	NNR3	HBDON
C4	M110	M110	M110	CHBAC	SO3	MR
D1	PSANEG	HUTTER	M119	IP	LIPIA	MR
D2	CRSURF	HUTTER	NROTB	QSUMN	XlogP	MR
D3	KHENI	M052	M006	GLOBUL	HALO	N/A
D4	MW	LIPIA	M110	QSUMPOS	NNR3	N/A
D5	MGHBA	M093	M092	MGHAC	MR	XlogP
D6	XlogP	HALO	NR3	CRSURF	XlogP	XlogP
D7	HUTTER	M048	M093	QSUMN	COOH	NRING
D8	BALESP	NROTB	FUNCGR	KHEPH	HBACC	NRING
E1	M074	M108	M108	SGEPH	NNR3	NRING
E2	COHESI	M003	GVW50	QSUMPOS	XlogP	HBDON
E3	KHESU	SlogP	M056	N/A	N/A	N/A
E4	HGHAC	NRING	M110	N/A	N/A	N/A
E5	QSUMNEG	HALO	NR3	N/A	ESTER	N/A
E6	T1E	NNR3	UNSUIT	N/A	HBDON	N/A
E7	HBDON	HALO	FUNCGR	N/A	XlogP	N/A
E8	QSUMPOS	M096	M128	N/A	HBDON	N/A
E9	M161	NR3	M164	DIPOLM	SO1	NRING
E10	M001	HALO	GVW80	HACSUR	XlogP	HBACC
E11	QSUMN	N/A	N/A	EHBBA	N/A	MR
E12	HACSUR	N/A	N/A	MGHBA	N/A	NROTB
E13	NROTB	M044	M044	RUGOS	SO2	NROTB
E14	PSANEG	HALO	M033	KHECA	NO2	NROTB
E15	HUTTER	M002	GVW50	QSUMN	HBDON	NROTB
E16	HUTTER	NO2	M067	T2E	NO2	XlogP

[a] Generated from all available descriptors. [b] Generated from 1D, 2D, and SMARTS-based descriptors. [c] Generated from 2D and SMARTS-based descriptors. [d] Generated only from 3D and quantum chemically derived descriptors. [e] Generated only from rapidly computable 1D and 2D descriptors. [f] Generated only from rapidly computable descriptors also used in druglikeness indices. [g] Shown in Figure 2. Descriptors are explained in Table 2.

below).[30] The value found for the molecular weight is, however, below the margin of 160 given by the same authors. Since this descriptor appears only in the third level, it is evaluated only for a small fraction of the compounds (1.5%) reaching that branching point and therefore represents a more specific criterion at that point. In turn this indicates that separating margins applicable to a wider range of compounds should be derived from early branching points in the decision tree. If descriptors appear at later branching points, then they are merely superseded by other available variables but conversely can be more predictive in the absence of the latter. The molecular weight is a typical example for that, since it is used as the very first criterion in the decision trees B, D, and F (see below) as well as in other studies.[30,49]

Although the molecular weight is immediately computable for a chemical compound, it does not provide much information about how druglike substances differ in their constitution from nondrugs. The same holds for the molecular volume and likewise the molar refractivity. To obtain utilizable information helpful to the synthetic design of pharmaceutical agents, the presence of certain chemical groups in drugs has been studied extensively.[5,23,26] Using decision trees, Wagener and van Geerestein yielded the occurrence of a hydroxyl group as the most important criterion, whereas Muegge and co-workers derived a lower limit of 1 and 2, respectively, functional groups that are expressed by so-called pharmacophoric points.[11,26] Decision tree A contains a number of descriptors related to the presence of such functional groups. M092 and M084 indicate the presence of secondary aliphatic amines and amides, respectively, that render a compound as potentially druglike. In turn the appearance of sulfonic acids (M110) is more common among nondrugs. Furthermore, Muegge et al. noted that nondrugs are typically underfunctionalized in comparison to pharmaceutical substances. Exceptions to this rule are psychoactive substances. As in the case of girisopam they often lack even an OH-group. To account for the total number of functional groups, the similar descriptor in this study, namely FUNCGR, comprises hydroxyl, amino, NO_2, sulfoxyl, sulfonyl, SO_3, nitrile, CF_3, CCl_3, and ester groups as well as a number of halogen atoms and carboxylic and other acids. Figure 3e shows the similar distribution among drugs and nondrugs. It can be seen that for 2 or more functional groups the presence of a pharmaceutical substance is more likely. This corroborates earlier results, especially the distribution of pharmacophoric points in the ACD and the MDDR as shown in the review of Muegge.[2,11,26] Since there is a substantial number of drugs (3.6%) without any functional groups among our data set (see Figure 3e), the decision tree put the similar separating margin to 1. The prominent class of drugs with very few functional groups found in this branch of the tree are steroids. They possess an extensive hydrocarbon skeleton that renders them rather featureless.

Compounds with 1 or more functional groups are then evaluated for the amount of their polar surface area (PSA) that has been recognized as limiting intestinal absorption and similarly blood-brain barrier permeability. Upper limits given in the literature are 120 Å,[2] especially for CNS-active compounds.[49,50] Although 34% of the drugs in our training set exhibit a larger PSA, less than 1% are found in this branch

Table 2. Explanation of Descriptors Used in the Decision Trees

descriptor	explanation
MW	molecular weight
XlogP	calculated XlogP[a 40]
SlogP	calculated SlogP[a 41]
MR	molar refractivity[a 41]
HBDON	number of hydrogen bond-donors
HBACC	number of hydrogen bond-acceptors
HALO	number of all halogen atoms
COOH	number of carboxylic acid groups
NOH	number of OH-groups
NO2	number of NO_2-groups
NNR3	number of protonatable NR_3-groups
SO1	number of S═O-groups
SO2	number of SO_2-groups
SO3	number of SO_3-groups
NACID	number of deprotonable acidic groups
CN	number of CN-groups
CF3	number of CF_3-groups
CCL3	number of CCl_3-groups
ESTER	number of ester-groups
FUNCGR	number of functional groups incl. NOH, ESTER, NO2, SO1, SO2, SO3, CN, CF3, CCL3, NNR3, and NACID
UNSUIT	number of unsuitable groups (according to ref 32 also including occurences of SMARTS given in Table S2 of the Supporting Information)
NRING	total number of rings
NR3	number of 3-membered rings
NROTB	number of rotatable bonds
LIPIA	number of violations of Lipinki's rule[b 28]
GVW80	compound qualified within 80% range of all drugs[b 30]
GVW50	compound in prefered range of 50% of all drugs[b 30]
OPREA	Oprea's critria for 70% of all drugs[b 29]
HUTTER	total Hutter's druglikeliness index[6]

SMARTS-derived descriptors:

M001	[OH1][#6,#7][!O]	hydroxyl, but not acid
M002	[N;H]	amines
M003	[S;H1]	sulfides
M006	[#8;X2]	−O−
M033	A1∼C∼A∼A∼A1	aliphatic 5-membered ring
M044	[#6]S(═O)(═O)[!O]	sulfonyl, but not sulfonic acid
M048	[#6]C(═O)[NH][C,H]	amido
M052	[CX4H3]	primary carbon
M056	C═[C;!CX2,C;!OX1,C;!NH2,C;!OH1]	alkenes, enes, enamines, and enols
M063	[C;H0]([OH1])([CX4])([CX4])[CX4]	tertiary alcohols
M067	cOc	diaryl ether
M074	CC(═O)Oc	alkyl-aryl ester
M084	[#6][NH0]([#6])C(═O)[#6]	amides
M092	C[NH1]C	secondary aliphatic amine
M093	CN(C)C	tertiary aliphatic amine
M110	[#6][SX4](═O)(═O)∼O	sulfonic acid
M119	C═[NX2][NX3;H1][!#7;!#8]	hydrazone
M128	[#6,#8][PX4;H0](═O)([OX2,#6,#7])[OX2]	phosphoric acid
M161	[A;r7]	aliphatic 7-membered ring
M164	[A;r10]	aliphatic 10-membered ring

quantum chemically derived descriptors:

MOLVOL	molecular van der Waals volume
GLOBUL	globularity[55]
DIPOLM	molecular dipole moment
IP	ionization potential
CHBAC	covalent hydrogen bond acidity[56]
EHBBA	electrostatic hydrogen bond basicity[56]
ESPMIN	minimum of molecular electrostatic potential (MEP)[57]
VMINUS	variance of negative MEP[57]
BALESP	balance parameter of the MEP[57]
QSUMN	sum of MEP derived atom centered point charges on nitrogen atoms[58]
QSUMNEG	sum of all MEP derived negative atom centered point charges[58]
QSUMPOS	sum of all MEP derived positive atom centered point charges[58]
PSA	polar surface area[50,59]
PSANEG	negative polar surface area (electrostatic potential < -25 kcal/mol)
CRSURF	ratio of surface on carbon atoms to the total surface area
HACSUR	ratio of the surface area belonging to atoms that are hydrogen bond acceptors to the total surface area
KHECA	Kier and Hall E-state on carbon atoms[60]

Table 2. (Continued)

quantum chemically derived descriptors:	
KHENI	Kier and Hall E-state on nitrogen atoms[60]
KHESU	Kier and Hall E-state on sulfur atoms[60]
KHEPH	Kier and Hall E-state on phosphorus atoms[60]
SGEPH	geometric E-state on phosphorus atoms[14]
G1GEO	gravitational G1 geometrical index[61]
T1E	topological electronic index 1[62]
T2E	topological electronic index 2[62]
MGHBA	minimal distance between two hydrogen bond acceptors
MGHAC	minimal distance between a hydrogen bond donor and an acceptor
RUGOS	rugosity (=molecular surface/molecular volume)[63]
COHESI	cohesive index[64]

[a] Additional parameters required for boron and silicon containing compounds developed in our laboratory are given in Table S4 of the Supporting Information. [b] Using the computed XlogP values.

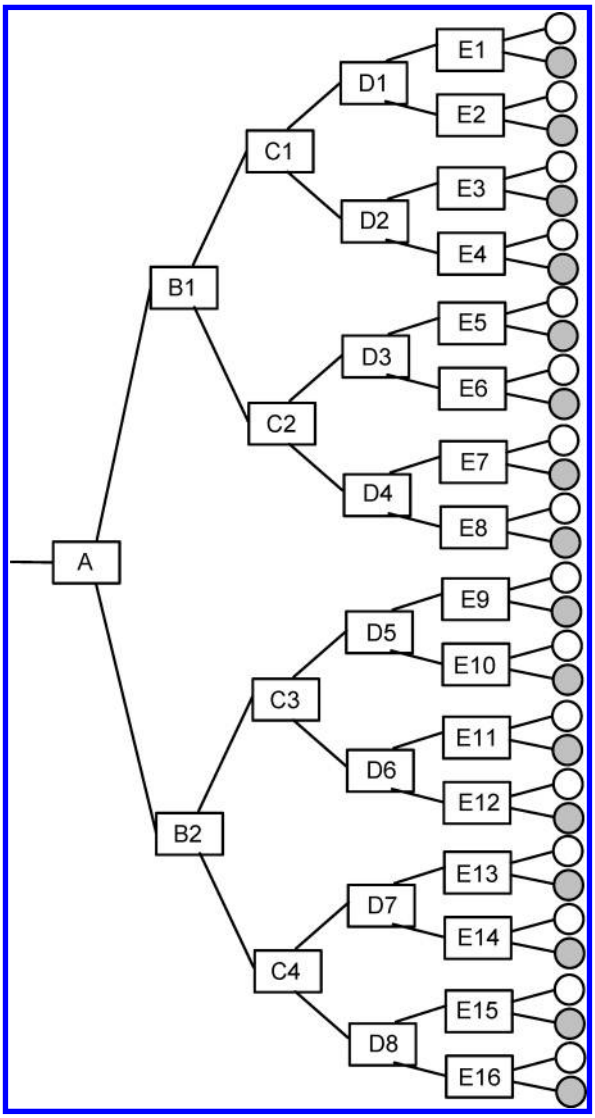

Figure 2. Topology of the obtained desision trees for separating drugs from nondrugs. Used descriptors are listed in Table 1 and briefly explained in Table 2.

of the tree where a separating margin of 102 Å^2 is set. Since PSA is superseded by other descriptors and not present in other trees, no recommendation regarding the polar surface area can be given at this point. Further descriptors appearing from the fourth level on that show up at earlier branching points in other decision trees and therefore apply to a considerable number of compounds will be discussed below. It is interesting to note that, except for Hutter's index, no other of the available druglikeness indices is present in this tree, even if the possible branching depth was increased.[6]

The predictive quality of the linear support vector machine trained with the same set of all descriptors (A) is not considerably better compared to the classification results of decision tree A (see Tables 3 and 4). Seemingly, the improvement regarding the prediction of drugs in the test set goes along with the worse result for the nondrugs. Wagener and van Geerestein noted a similar trend for decision trees.[11] Despite the similar prediction accuracy achieved by the decision tree and by the support vector machine, there is only moderate overlap regarding the classification of individual compounds. The correlation is highest for the 51 compounds comprising the set of new drugs ($r = 0.693$) and lowest for the drugs of the training set ($r = 0.345$). These differences in classification between decision tree and support vector machine were more closely analyzed for the 536 compounds of the test set, with respect to their pharmaceutical categories. Considered were those categories that contained at least 4 compounds in the test set. Figure 4 shows the percentage of falsely predicted compounds in these categories. Apparent is the failure of both classification algorithms for acaricides and herbicides. These nondrugs are consequently misclassified. This means, however, that acaricides as well as herbicides contain features that distinguish them from ordinary chemicals and render them as being druglike. Considering the biological functions of these agents that imply mostly enzyme inhibition, this is, however, not an actual misclassification after all. The situation is similar for anthelmintics that exhibit a toxicological function onto parasites. Interestingly, the fraction of misclassified drugs among the antibacterial penicillines and antivirals is much lower. Other drugs that are frequently misclassified act on the central nervous system, e.g., nootropics, CNS-stimulants, sedatives, and hypnotics. These substances have to pass the blood-brain barrier and therefore are usually more lipophilic than other drugs.[30] A further untypical category of drugs is local anesthetics that are nonorally administered in contrast to the majority of drugs. No misclassifications were observed among ACE-inhibitors, narcotic analgesics, antianginals, antihypertensives, antimigraines, nonsteroidal anti-inflammatories, antipsychotics, antineoplastics, macrolide antibiotics, and antiamebics. The 4 latter categories are in turn no typical drugs in the sense of being orally administered and acting on peripheral human targets. Thus no general conclusion about the prediction accuracy for typical and nontypical drugs can be made.

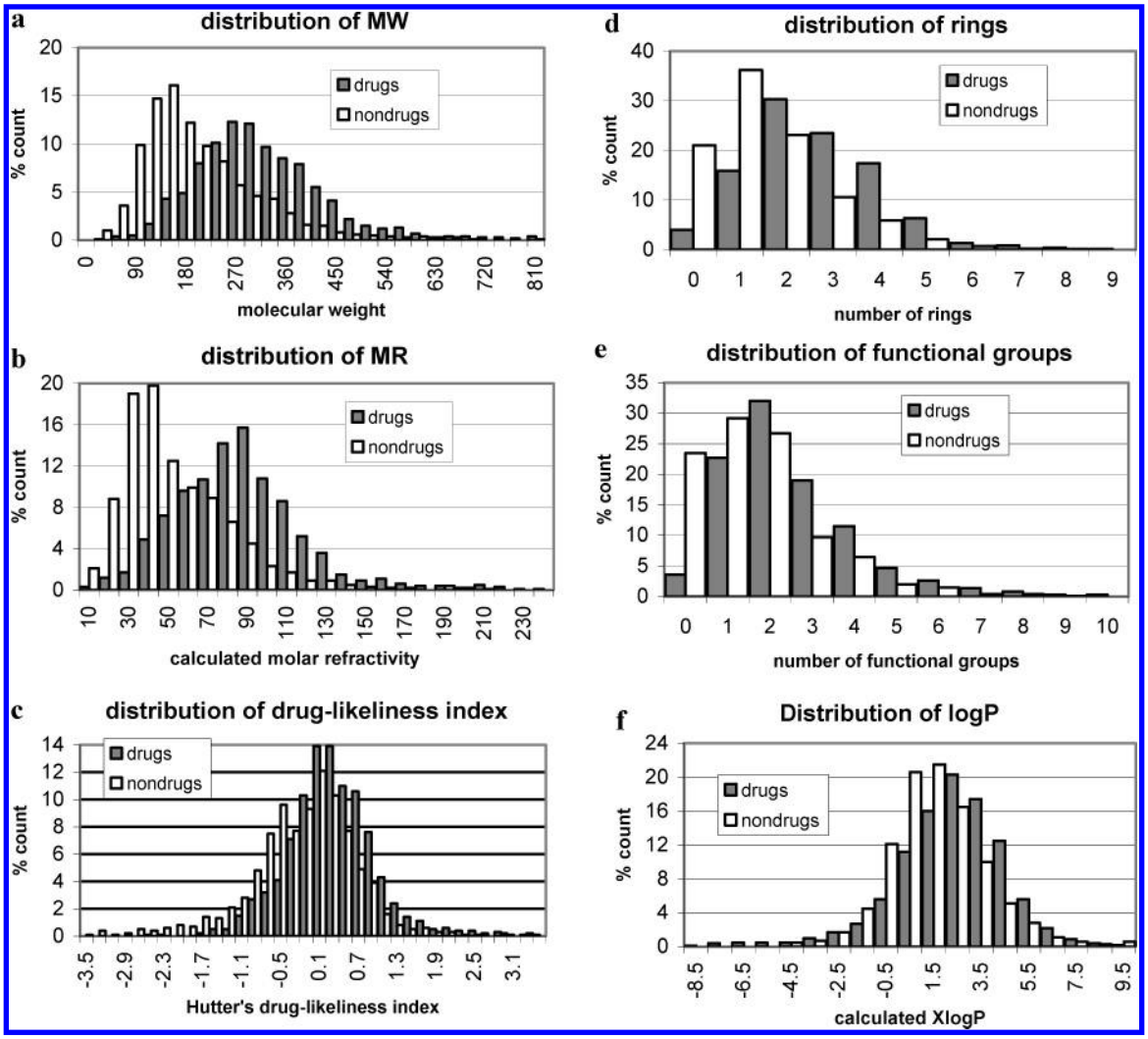

Figure 3. Distribution of single descriptors among the drugs and nondrugs in the data set: a) molecular weight, b) molar refractivity, c) Hutter's druglikeliness index, d) number of rings, e) number of functional groups, and f) calculated XlogP.

Table 3. Obtained Prediction Accuracy of the Decision Trees[c]

		after decision tree					
compound set	no. of compds	A	B	C	D	E	F
drugs training	2805	90.1	89.5	87.5	86.8	86.8	81.3
nondrugs training	2014	77.5	76.3	74.3	77.3	76.3	77.8
drugs test	312	92.6	95.5	91.4	86.2	90.7	87.2
nondrugs test	224	32.6	30.4	37.5	35.7	34.4	38.8
new drugs	51	92.2	96.1	88.2	82.4	94.1	90.2
drugs test MS[a]	110	97.3	93.6	94.5	95.5	91.8	85.5
nondrugs test MS[a]	49	44.9	30.6	24.5	46.9	28.6	32.7
drugs valid MS[a]	61	98.4	93.4	93.4	93.4	91.8	83.6
nondrugs valid MS[a]	18	44.4	55.6	38.9	50.0	50.0	66.7
no. of available descriptors	–	249	185	163	64	26	8
no. of actually used descriptors	–	28	22	20	22	15	8
ratio used/available[b]	–	11.2	11.9	12.3	34.4	57.7	100

[a] Compounds taken from Murcia-Soler et al.[9] [b] Given in percent. [c] Given in percent.

Table 4. Obtained Prediction Accuracy of the Support Vector Machines[b]

		after support vector machine					
compound set	no. of compds	A	B	C	D	E	F
drugs training	2805	90.7	89.3	88.2	88.4	88.8	80.4
nondrugs training	2014	80.0	88.0	87.5	79.5	73.4	75.6
drugs test	312	96.5	94.2	95.8	90.4	91.0	87.2
nondrugs test	224	27.7	29.9	31.3	34.8	33.9	33.0
new drugs	51	96.1	90.2	92.2	88.2	94.1	90.2
drugs test MS[a]	110	94.5	90.0	90.9	93.6	94.5	85.5
nondrugs test MS[a]	49	65.3	75.5	73.5	71.4	61.2	55.1
drugs valid MS[a]	61	91.8	90.2	86.9	90.2	91.8	77.0
nondrugs valid MS[a]	18	50.0	72.2	66.7	50.0	55.6	61.1
no. of available descriptors	–	249	185	163	64	26	8

[a] Compounds taken from Murcia-Soler et al.[9] [b] Given in percent.

Figure 4 reveals furthermore the different classification results between decision tree and support vector machine. In many cases false predictions among a pharmaceutical category are only present in either decision tree or support vector machine, whereas in the remaining categories the percentage of false classification is comparable between both algorithms.

Both decision tree A and support vector machine A were trained with the full set of available descriptors. Due to the presence of computationally more demanding descriptors such as the molecular volume, the polar surface area, and SMARTS-based descriptors, the question arises if it is possible to omit such descriptors without worsening the performance of the prediction algorithms considerably. Moreover, a reduction of the number of available variables

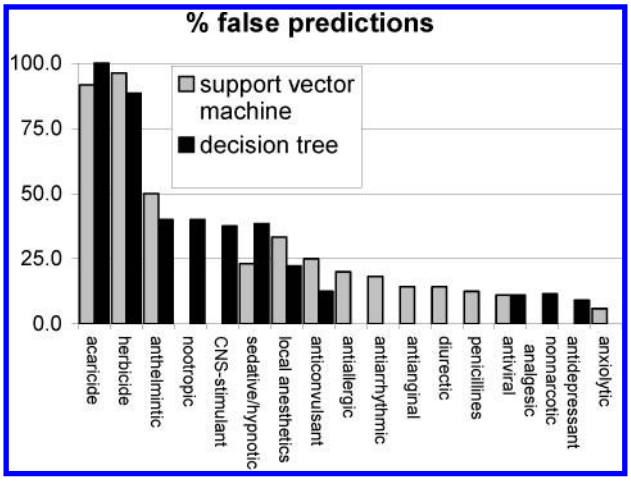

Figure 4. Comparison of falsely predicted compounds from decision tree A and support vector machine A, according to their pharmaceutical category.

reduces the dimensionality of the fitting problem and would allow a more general applicability of the derived classification scheme. Therefore, we deduced subsets of the complete set of descriptors. In set B the quantum chemically derived variables were removed, leaving 185 descriptors that comprised only 1D, 2D, and SMARTS-based variables. In addition, even more reduced subsets were also investigated.

Comparison of the results obtained by tree B to those of tree A shows that the prediction accuracy is not worsened considerably due to the renunciation of computationally expensive descriptors (see Table 3). The first descriptor appearing in tree B is the molecular weight replacing the molecular volume to which it is highly correlated. As separating margin a value of 230 was obtained. The comparison between drugs and nondrugs in our data set shows that the respective distributions of the molecular weight intersect around this value (see Figure 3a) with a clear tendency of drugs toward higher values. Similar trends between databases containing drugs, i.e., the MACCS-I Drug Data Report (MDDR) as well as the Comprehensive Medicinal Chemistry (CMC) and compound collections comprising mainly nondrugs such as the Available Chemical Directory (ACD), were reported before by Oprea and later by Zheng et al.[29,51] While Figure 3a shows a clear separation of the maximum peaks of the molecular weight of nondrugs and drugs, Zernov et al. yielded a stronger overlap between both distributions that intersect at around 360 that is higher than our separating margin.[19] Almost complete overlap between substances from the World Drugs Index (WDI) and those from the ACD were found by Li et al.[17]

Another obvious visual separation is found for the molar refractivity (see Figure 3b). The separating margin obtained by tree B is 40, identical to that obtained from tree A. Although such an obvious gap between the peaks of the distributions of drugs and nondrugs is not visible for Hutter's druglikeliness score (see Figure 3c), this index is found in the second level of tree B and tree E, emphasizing its importance. According to the underlying statistical derivation, compounds exhibiting a value above zero are more likely to be drugs. Still, 22.6% of the considered drugs showed negative values.[6] The separating margin derived by tree B (−1.05) captures in particular steroids that are typically scored too low by this index due to their featureless nature.

There are some strong similarities regarding certain descriptors in tree A and B. M092 (secondary aliphatic amines), MR (molar refractivity), and M110 (sulfonic acids) are found at the same branching points with identical margins. Since they furthermore appear in the early levels, they seem to be of particular importance. In contrast to tree A, the number of rotatable bonds (NROTB) and the number of violations to Lipinski's rule (LIPIA) appear for the first time. According to their margins, drugs should possess at least one rotatable bond and should not show any violations to Lipinski's rule.[28] Due to their presence in the fourth level, the derived margins may, however, not be generally applicable. Recently Oprea and co-workers showed that 90% of marketed drugs and those in clinical trails have up to 2 violations and up to 21 rotatable bonds.[52] The preferred range is between 2 and 8 bonds.[29] Oprea's definition of nonterminal rotatable bonds, however, accounts for the partial flexibility of nonaromatic 6-membered and larger rings.[29,52] E.g. a cyclohexane ring is assigned two rotatable bonds and a cyclohexene ring one rotatable bond. In contrast we consider only such nonterminal bonds as freely rotatable, that are not in a ring. This explains that we find a lower number of rotatable bonds as margin using a more stringent definition.

To elucidate the relevance of substructure-based descriptors and common druglikeness indices, the available variables for the following tree C were further restricted, leaving out all 1D descriptors and Hutter's druglikeliness index. Here, we found the number of rings (NRING) at the first branching point with a separating margin of 1. Within our data set the number of rings is strongly correlated to the number of 6-membered rings ($r = 0.898$) that comprise the most frequently found ones among all ring systems.[22] The distribution of rings is shown in Figure 3d. 96% of the considered drugs in the data set contain at least one or more rings, while conversely 21% of the nondrugs possesses no cyclic structure. Statistics of larger drug databases revealed very similar distributions (98% of compounds in the MDDR and 95% in the CMC, respectively), whereas the ACD contains less acyclic nondrugs (about 7 to 11%) compared to our data set.[29,51] Again M092 and M110 are found at identical branching points as in tree A and tree B. Likewise, M093 (tertiary aliphatic amine) is adopted from tree B at C1. Occurrence of a 7-membered ring (M161) also renders a compound as being druglike since they are frequently present.[22] The most prominent representatives of such rings are substituted benzodiazepines (e.g., diazepam and imipramine). These comprise 6.9% of the drugs in the training set, whereas only 0.3% of the nondrugs contain a 7-membered ring. Despite the presence of NR3 and M164 that indicate 3- and 10-membered rings, respectively, no accordingly clear separation was found, since these structural differences are evaluated only for a small fraction of the data set (less than 2%). For the number of rotatable bond two different margins were found (1 and 2, respectively) at branching points B2 and C2. In this context it should be noted that from the indices indicating druglike compounds, those by derived by Ghose and co-workers (GVW80 and GVW50) are used only at later stages (fifth level).[30]

Other descriptors related to the molecular topology have been studied by Li et al.[17] They applied extended connectivity fingerprints ECFP_4 that characterize the vicinity around a heavy atom up to 4 bonds. This allowed the detection of

Table 5. Obtained Results after Applying Decision Trees and Support Vector Machines, Respectively, Successively[c]

		after decision tree			after support vector machine		
compound set	no. of compds	E → A	E → C	E → C → D	E → A	E → C	E → C → D
drugs training[a]	2805	83.8	81.3	77.6	86.3	84.1	81.0
nondrugs training[b]	2014	82.9	85.3	89.6	86.1	90.5	92.9
drugs test[a]	312	87.5	84.9	76.3	90.1	89.4	83.7
nondrugs test[b]	224	41.5	48.2	55.4	39.7	44.6	51.3
new drugs[a]	51	90.2	86.3	74.5	90.2	88.2	82.4

[a] Total percentage of drugs retained after applying the classification schemes successively. [b] Total percentage of nondrugs filtered out after applying the classification schemes successively. [c] Given in percent.

substructures that are statistically associated with their drug and nondrug data sets, respectively. These features do, however, enable a classification only in the context of the applied support vector machine approach and are moreover larger than usual functional groups. In order to find similar topological indices to classify druglike compounds, Gálvez et al. applied linear discriminant analysis.[12] They yielded an equation comprising 6 descriptors that produced a prediction accuracy of 82% for 2000 drugs taken from the Merck Index.[21] These indices reflect molecular fragments, however, only indirectly and are more complex than the substructure patterns and SMARTS strings used here. Therefore no simple guidelines for the design of druglike substances can be derived using these topological indices.

Since quantum chemically derived quantities comprise a similar complex kind of variables, we trained decision tree D, and likewise support vector machine D, exclusively with such descriptors. Like in tree A the molecular volume is used as first branching criterion. It is followed by a number of descriptors that are derived from the molecular electrostatic potential and are related to the charge distribution on the molecular surface. PSANEG is the negative polar surface area and corresponds to the contribution of electronegative atoms to the total surface area. As a consequence, compounds with a negative polar surface area larger than 39 Å^2 are more likely to be nondrugs. In a qualitative fashion this corresponds to the limiting number of nitrogen and oxygen atoms in Lipinski's rule and Oprea's criterion.[29] At first sight, it seems to be rather difficult to find a similar counterpart to VMINUS, the variance of negative electrostatic potential. At the same branching point in tree A (B2, see Table 1) we find FUNCGR, the number of functional groups. Substitution with functional groups introduces mostly electronegative elements to the molecular framework and thus increases the variance among the negative electrostatic potential, particularly upon substitution with different functional groups. Here, the kind of electronegative atoms seems to be decisive as expressed by ESPMIN, the minimum of the electrostatic potential, and likewise by CHBAC, the covalent hydrogen bond acidity. A similar qualitative counterpart is M110 from tree A and tree B that accounts for the occurrence of sulfonic acids that are more frequently found in nondrugs, in contrast to hydroxyl groups, for example. For compounds containing phosphorus, another electronegative element, a similar separation is more difficult, since derivatives of phosphoric acid appear in drugs as well as in nondrugs. Nevertheless, the kind of substitution seems to be decisive, as KHEPH, the Kier and Hall E-state on phosphorus atoms, is chosen as separator, reflecting their chemical vicinity. Corresponding E-states for carbon, nitrogen, and sulfur also appear in tree D and tree A but at rather late branching points. From the remaining descriptors used as branching criteria, no additional unequivocal information regarding the design of similar compounds could be derived.

The completely contrary approach to tree D regarding descriptors was pursued using set E. Here, only rapidly computable 1D and 2D descriptors and indices are applied. Like in tree B the molecular weight is the first criterion. In the second level Hutter's index is used twice with two different margins of 0.77 at B1 and −1.05 at B2, respectively, as in tree B. The first margin accounts for nondrugs that exhibit particularly high scores. Nondrugs with a score between 0 and the margin at 0.77 comprise about 36% of all nondrugs.[6] The molar refractivity is found at the same branching point C2 as in tree B with the identical margin (40) and again at D5 with a lower margin of 30 for the remaining drugs. As expected M110 is replaced by SO3 that holds similar information about the occurrence of sulfonic acids and derivatives, since both variables are strongly correlated ($r = 0.922$). Although the two methods for calculating the water/*n*-octanol partitioning coefficient logP, XlogP, and SlogP are also correlated ($r = 0.753$), XlogP is used exclusively. The obtained separating margins, however, range from −3.4 to +10.6 depending on the considered subset of compounds. The usual range of logP found in drugs is −0.4 to +5.6, whereby substances targeting the central nervous system tend to be even more lipophilic.[30] Moreover, XlogP is found only at later branching points, reflecting classifications based on subtle differences in the absence of other more decisive descriptors. Thus no general recommendation regarding the logP range of potential drugs can be given here.

Since not all of the 5 criteria accounting for druglike substances appear in the decision trees A to E, it is of interest to elucidate their pairwise dependence. As can be seen from Table 6, the intercorrelation between these indices is rather low. The largest value ($r = 0.377$) is found between the 50% and the 80% druglike range of Ghose et al.[30] Surprisingly, compounds fulfilling Lipinski's rule show a slightly negative correlation to actual drugs and Hutter's index, respectively. This is due to the fact that the majority of nondrugs also fulfill these criteria for oral bioavailability (83.6% in the data set), while conversely there are numerous compounds among the drugs that are nonorally administered, e.g., antineoplastics, or that violate Lipinski's rule (29.3% of the drugs in the data set). These results confirm the findings of Oprea who did not find a significant difference between drugs and nondrugs based on the descriptors used in Lipinski's rule.[29] It is therefore interesting to note that the descriptor related to Lipinski's rule (LIPIA) appears anyhow in trees B and E,

Table 6. Intercorrelation between Indices Indicating Druglike Substances and Actual Drugs in the Data Set

	LIPR[a]	OPREA[b]	GVW80[c]	GVW50[d]	HUTTER[e]	ISDRUG[f]
LIPR	1.000	0.117	0.242	0.186	−0.032	−0.156
OPREA	0.117	1.000	0.261	0.154	0.184	0.183
GVW80	0.242	0.261	1.000	0.377	0.157	0.331
GVW50	0.186	0.154	0.377	1.000	0.116	0.225
HUTTER	−0.032	0.184	0.157	0.116	1.000	0.282
ISDRUG	−0.156	0.183	0.331	0.225	0.282	1.000

[a] No violation of Lipinski's rule.[28] [b] Within 70% of druglike compounds according to Oprea.[29] [c] Within the 80% range of all drugs according to Ghose et al.[30] [d] Within the preferred 50% range of all drugs according to Ghose et al.[30] [e] Druglike according to Hutter's index.[6] [f] Being an actual drug.

whereas Oprea's criterion is not used (see Table 1). Comparison of the separating margin (that is 1) with the distribution of compounds in the data set revealed that from all drugs 70.7% fulfill Lipinski's rule (molecular weight not above 500, logP below +5, less than 5 hydrogen-bond donors, and less than 10 hydrogen-bond acceptors), 19.9% show precisely 1 violation, 6.8% exhibit 2, and only 2.5% contain 3 violations, respectively. This distribution is quite similar to that found by Oprea et al. for a much larger data set, stating that 90% of the marketed drugs and those in clinical trails have up to 2 violations, whereas the more recent compounds contain more violations.[52] This is attributed to higher molecular weight and more hydrogen-bond acceptors. The same trend also transpires among our set of new drugs. Here, only 52.9% match Lipinski's rule, 23.5% show 1, 16.6% show 2, and 5.8% contain 3 violations, respectively. Conversely, only 3.6% of the nondrugs in our data set show 2 or more violations. If not before, one has to recall that Lipinski's rule was derived as criterion for bioavailability in the first place, rather than for druglikeness.

Since most of the druglikeness indices were only used at later stages in the decision trees so far, indicating that these criteria are superseded by other descriptors, it is obvious to test the performance of the classification algorithms if only such variables are available that have been used to derive the corresponding indices. Set F thus contains only 8 descriptors, namely molecular weight, XlogP, and molar refractivity as well as the respective count of rings, rotatable bonds, functional groups, hydrogen-bond donors, and hydrogen-bond acceptors. Like in trees B and E, the first branching criterion chosen is the molecular weight. XlogP and the number of functional groups comprise the second level, separating particular hydrophilic compounds and those without any functional groups. Again the separating margins determined for XlogP at the various branching points cover a large range and do not allow a general recommendation. For the molar refractivity, the margins indicating druglike compounds are somewhat higher (56, 62, and 75, respectively) than in trees A, B, and E but closer to the lower limit of 70 given for the preferred range of drugs according to Ghose et al.[30] The count of rotatable bonds is used to separate nondrugs from the fraction of drugs that exhibit a particularly high number of freely rotatable bonds. Within our data set we found that 15% of all drugs possess more than 12 rotatable bonds according to the more stringent definition used here. The margins found for the number of rings suggests that drugs should have 1 or more rings, in accordance with the results determined by trees B and C. The count of hydrogen-bond donors and hydrogen-bond acceptors is once again applied to a small fraction of the compounds indicating that drugs should contain at least one of these features arising from nitrogen or oxygen atoms. Since these are also involved in most functional groups, Wagener and van Geerestein noted their importance particularly in alcohols and tertiary and secondary aliphatic amines as well as in the hydroxyl groups of phenols, enols, and carboxyl groups.[11] Corresponding descriptors representing secondary and tertiary aliphatic amines, namely M092, M093, and NNR3, respectively, are frequently found in our decision trees (see Table 1). Since we capture isolated hydroxyl groups separately from those being part of carboxylates, only the count of the latter (COOH) turns up in tree E.

Identifying properties that render compounds as nondrugs is conversely a rather difficult task.[11] Either these remain featureless, for example pure hydrocarbons, after substances with similar druglike features have been separated or contain rather specific and complex functional groups, such as those represented by the SMARTS string for unsuitable compounds. The similar descriptor UNSUIT is, however, found only in tree C in the last level. Analysis of the data set showed that 25.9% of all drugs contain 1 or more of these unsuitable features, whereas conversely 65.1% of the nondrugs do not contain any of these features. Thus the occurrence of unsuitable fragments cannot be used to classify substances as nondrugs right away, except for those fragments that render a compound as reactive or toxic in general.

Despite the different underlying descriptor sets, the decision trees show comparable results (see Table 3). As expected, the prediction accuracy for drugs is highest for tree A that comprises all descriptors and decreases upon reducing the number of available variables. It is interesting to note that in trees A to E only a fraction of the available descriptors is actually used to perform the classification, thus reducing the dimensionality and likewise the chance of overfitting. In tree F, on the other hand, the number of available variables is obviously not large enough to obtain a better classification and to populate the branching points sufficiently. In this context, tree E shows the most balanced result regarding prediction accuracies with respect to all compound sets. Moreover, all of the applied descriptors can be rapidly computed from the molecular structure and thus used for filtering large compound libraries. Alternatively, tree B can be used because it contains only 11 SMARTS based queries on top of the range of descriptor computed for tree E, while keeping more actual drugs than that.

Quite similar results regarding the classification behavior of decision trees were reported by Wagener and van Geerestein.[11] They also found that an improved prediction accuracy for drugs (fewer false negatives) goes along with an increased number of misclassified nondrugs (more false positives). We observed this trend in particular for the classification results comparing training and test sets for decision trees as well as support vector machines (see Tables 3 and 4). Although support vector machines are known to yield very good results in performing classification tasks, the obtained accuracies are not throughout better than those of the decision trees as observed in a related study but rather similar.[20] Müller and co-workers compared a number of machine learning methods but used nonlinear kernel func-

tions (radial basis and polynomial, respectively) for the support vector machines instead of the linear function applied here.[20] Recently, Li et al. applied a probabilistic support vector machine with a radial basis that yielded better results than with a linear kernel, as expected.[17] While the linear function corresponds to a separating line in a 2-dimensional descriptor space, radial basis and polynomial functions introduce curvatures. It is easy to see that such nonlinear kernel functions render support vectors machines susceptible to overfitting, e.g., by trying to account for outliers in the data set. Thus the more robust linear kernel function was used here to obtain a more generally applicable classification scheme. Moreover, a linear kernel function closely resembles the decision function (the separating margin) used in the decision trees. Since both decision trees and support vector machines optimize the separating margins of the available descriptor space, they are less sensitive toward extreme data compared to neural networks that cannot make reliable predictions about compounds containing numerical values outside the trained range. Thus, decision trees and support vectors machines are less affected by the actual feature distribution in the training set than neural networks that are well-known to be susceptible to overtraining.

The large difference in the prediction accuracies for the nondrugs between the training and the test set is somewhat surprising as the splitting of the data set was carried out on the basis of a cluster analysis that is expected to yield an adequate sample that is more representative than a random selection. A possible explanation is that nondrugs are more evenly dispersed in chemical space. Conversely, drugs form more emphasized clusters as it has been postulated earlier.[53,54] As a consequence the centroids of the clusters that are used as molecules in the test set are less representative in the case of nondrugs. Conversely, the obtained accuracies for the set of new drugs are quite similar to those of the other drug sets, supporting this reasoning. Good and Mermsmeier have investigated the effect of different test set selections in the context of druglikeness classification.[16] For drugs they compared randomly chosen test sets to a number of sets that were derived from drug ontology classes. Since certain substructures and substituents are, however, not evenly distributed among the drug ontology classes, we tried to represent the chemical diversity using a cluster analysis, instead. This also circumvents the problem of generating similar classes and ontologies for nondrugs, respectively.

For a further comparison we used the data set of Murcia-Soler et al., since they list the similar compounds individually.[9] Moreover, they partitioned their data set accounting for a similar fractional distribution of the drugs from a total of 11 therapeutic categories among the training and test sets. Similar results obtained with our decision trees and support vector machines using their test and validation sets are given in Tables 3 and 4. Mucia-Soler et al. applied a neural network technique whereby 62 topological indices were used as the input layer. This kind of descriptor selections resembles most closely the 1D and 2D descriptors available in decision tree E and in support vector machine E. For the drugs of the test and validation sets we obtained prediction accuracies of 91.8% or better from both approaches compared to the reported 76.4% and 77.1%, respectively. Conversely, the neural network approach yielded better results for the nondrugs (70.2% and 76.0%).[9] Here, we observed a superior performance of the support vector machine E (61.2% and 55.6%) in comparison to the decision tree E (28.6% and 50.0%). This is again in line with the findings of Wagener and van Geerestein on the performance differences of decision trees regarding drugs and nondrugs.[11] Seemingly, it is more difficult to identify nondrugs than drugs using decision trees. This is also reflected by the larger differences of the prediction accuracies obtained for the nondrugs (28.6%−55.6%) compared to the drugs (83.6%−98.4%). Despite the partitioning criteria regarding the compound selection for the drugs comprising the test and validation sets were different (similar percentage of each therapeutic category) compared to our test set (similar distribution of features according to a cluster analysis), the performance of the decision trees as well as of the support vector machines is comparable for all sets used for testing. In particular for decision tree E and support vector machine E where only descriptors were used that are comparable to the topological indices of Murcia-Soler and co-workers, the prediction accuracy is even slightly higher for the drug test and the validation set (>91.8%) compared to our substantially larger test set (>90.7%) that furthermore also contains drugs from other therapeutic classes than the 11 categories. Thus, the chemical space of drugs is obviously quite similar among the various drug classes, and therefore the choice of compound selection for the test set has less influence on the results. In contrast nondrugs are chemically more diverse and thus more difficult to represent adequately.[53,54] In turn the better performance of the neural network approach of Murcia-Soler et al. may also be a consequence of the manageable number of compounds used for training and testing (417, 177, and 86, respectively) in conjunction with the high number of descriptors (62) for the input layer, leading to a compounds per descriptor ratio of 6.7. For comparison: the original study of Sadowski and Kubinyi had 108.7 compounds per descriptor.[3]

Due to the much larger number of nondrugs being recognized as drugs (false positives) and vice versa, the question arises whether it is feasible to apply classification schemes based on different descriptors sets successively to sort out more of the nondrugs. In the Introduction we have mentioned a corresponding strategy whereby the whole set of compounds is at first subject to a filter based on descriptors that can be computed rapidly. In the second and possibly a third step descriptors will be necessary that are derived from more demanding substructure queries and eventually from quantum chemical calculations.

Table 5 summarizes the results of this strategy depicted in the right-hand side of Figure 1. Consequently, the classification algorithms based on the rapidly computable descriptor set E comprise the first filtering step. For the second step, either descriptor set A (all descriptors) or set C (2D and SMARTS based queries) is eligible. Since quantum chemically derived descriptors are already present in set A, only the sequence E, C, and D is reasonable for a three-step filtering. Comparison of the results between the two algorithms shows that the support vector machines retain throughout more drugs than decision trees (see Table 5). Conversely, the decision trees achieve better results regarding nondrugs in the test set. Moreover, when applying support vector machines, all descriptors of the respective set have to be computed, whereas decision trees require only the

specified variables (see Table 1). This is a clear advantage if computationally demanding variables are necessary, e.g., for the partitioning schemes A, C, and D.

As a general trend we find that classifiers are selected frequently by a decision tree which do not show an obvious difference between drugs and nondrugs when inspecting their distributions visually such as shown in Figure 3 for the count of functional groups, Hutter's index, or the XlogP (Figure 3f). Moreover, the determined separating margins for such descriptors are often not identical with the numerical value one would derive manually. For example, the distribution curves of the molar refractivity (Figure 3b) intersect at 60, whereas decision tree B puts the separating margin at 40. This highlights one of the strengths of decision trees in deriving classification rules.

As expected, the accuracy of the decision trees is higher for all data sets compared to solely using Hutter's druglikeliness index.[6] Applying only this statistically derived criterion yields accuracies for drugs and nondrugs of 74.8% and 52.8%, respectively, for the training set (79.5%) and the test set (35.3%), respectively, and 82.4% for the set of new drugs. Due to the best value found in the last set, the question arises whether these recent drugs are more druglike than older ones. We therefore calculated the average of Hutter's score for the drugs in the training set (0.408), the test set (0.539), and the set of new drugs (0.507). Due to the large standard deviations (0.962, 0.825, and 0.488, respectively) and the small number of substances in the set of new drugs (51) compared to that in the training and test sets, however, the only reasonable statement that can be made is that these new drugs tend toward higher scores. Moreover, the prediction accuracy for these substances obtained by the decision trees and support vector machines is higher than that for the drugs in the training and test sets, except when using descriptor set D.

An and Wang have used a machine learning algorithm that is related to Hutter's druglikeliness index for performing classification of potentially active compounds.[15] The so-called ELEM2 method uses attribute-value pairs to iteratively generate rules that allow a complete separation among the training set. Compounds are assigned to a class on the basis of a ranking score that accounts for the differential likelihood of belonging to a class according to the derived set of rules. Except for decision trees, this algorithm achieved better results than *k*-nearest neighbor and neural networks.[15]

Two of the pharmaceutical compounds from the set of new drugs are classified wrong by all 6 decision trees (see Table 7): SGS-742 (3-aminopropylbutylphosphinic acid) CAS no. [145537-81-1] and YKP-509 ([(2R)-2-(2-chlorophenyl)-2-hydroxyethyl]carbamate CAS no. [194085-75-1]) possess both a molecular weight below the margin of 230 and are targeting the central nervous system. The former is an acyclic phosphine that is more likely to be found among nondrugs as reflected by its druglikeliness score (−0.10). Conversely, YKP-509 possesses a carbamate group that is exclusively found among the drugs in our data set resulting in an emphasized positive score of +0.83 (see Table 7) since all other structural features resemble those of drugs. Nevertheless both compounds do not contain a secondary or tertiary aliphatic amine. In combination with their low molecular weight this leads to the misclassification as nondrugs.

Table 7. Results for the 51 Pharmaceutical Compounds That Were Used as External Validation Set

	classification using decision tree[a]						
name	A	B	C	D	E	F	Hutter index[b]
ABT-510	+	+	+	−	+	+	1.37
ambrisentan	+	+	+	+	+	+	0.19
AMG-706	+	+	+	+	+	+	0.31
apricitabine	+	+	+	+	+	+	1.78
armodafinil	+	+	+	+	+	+	0.91
asoprisnil	+	+	+	+	+	+	−0.17
bicifadine	+	+	+	−	+	−	1.74
bifeprunox	+	+	+	+	+	+	0.19
canertinib	+	+	+	+	+	+	0.10
ceftobiprole	+	+	+	+	+	+	1.21
celegosivir	+	+	+	+	+	+	−0.12
chloroscoulerine	+	+	+	+	+	+	0.42
dabigatran	+	+	+	+	+	+	0.62
ED-71	+	+	+	+	+	+	−0.24
elocalcitol	+	+	+	−	+	+	0.05
eltrombopag	+	+	+	+	+	+	0.24
etravirine	+	+	−	+	+	+	−0.02
fospropofol	−	+	−	−	+	+	0.26
ibutamoren	+	+	+	+	+	+	0.92
indacaterol	+	+	+	+	+	+	0.64
isatoribine	+	+	+	+	+	+	0.94
ispinesib	+	+	+	+	+	+	0.62
lapatinib	+	+	+	−	+	+	1.07
LBM-415	+	+	+	+	+	+	1.24
maraviroc	+	+	+	+	+	+	0.75
nelarabine	+	+	+	+	+	+	0.50
paliperidone	+	+	+	+	+	+	0.64
pazopanib	+	+	+	+	+	+	0.26
prinaberel	+	+	+	−	+	+	0.44
Relacatibe	+	+	+	+	+	+	1.06
retapamulin	+	+	+	+	+	+	−0.59
rimonabant	−	+	−	+	+	+	0.84
rivaroxaban	+	+	+	+	+	+	1.10
sarizotan	+	+	+	+	+	+	0.17
satavaptan	+	+	+	+	+	+	0.27
seletracetam	+	+	+	+	+	−	2.68
SGS-742	−	−	−	−	−	−	−0.10
sitagliptin	+	+	+	+	+	+	1.01
sunitinib	+	+	+	+	+	+	0.36
tafluprost	+	+	+	+	+	+	0.43
taltobulin	+	+	+	+	+	+	0.70
tamibarotene	+	+	+	+	+	+	−0.05
tapentadol	+	+	+	+	−	−	−0.05
tebipenenem-pivoxil	+	+	+	+	+	+	0.98
teriflunomide	+	+	−	+	+	+	0.45
terutroban	+	+	+	+	+	+	0.74
torcetrapib	+	+	+	−	+	+	0.98
udenafil	+	+	+	+	+	+	0.43
valopicitabine	+	+	+	+	+	+	0.54
vandetanib	+	+	+	+	+	+	−0.13
YKP-509	−	−	−	−	−	−	0.83

[a] Compounds classified as drugs are denoted with a "+" sign. [b] Druglikeliness score computed according to ref 6. A positive value indicates a druglike compound.

This leads us back to the initial question as to which features render a compound being druglike. In contrast to studies where existing substance databases were analyzed in terms of features distribution, we applied decision trees to derive similar margins for important descriptors that can be used to separate drugs from nondrugs.[19,29,30,51,52] The following descriptors were found at early branching points whereby the majority of substances can be classified correctly: the molecular weight, the molar refractivity, Hutter's druglikeliness index, the number of ring systems, and rotatable bonds as well as the presence of secondary or

Table 8. Comparison of Rules That Indicate a Potential Druglikeness of a Compound

descriptor	this study	Oprea[a]	Ghose[b]	Xu[c]	Walters[d]
molecular weight	>230	*e*	160−480	*e*	200−500
H-bond donors	>0	≤2	*e*	≤5	≤5
H-bond acceptors	>0	2−9	*e*	≤10	≤10
logP	*e*	*e*	−0.4−5.6	*e*	−2.0−5.0
rotatable bonds	>0	2−8[f]	*e*	3−35	0−8
number of rings	>0	1−4	*e*	1−7	*e*
molar refractivity	>40	*e*	40−130	*e*	40−130
number of atoms	*e*	*e*	20−70	10−50[g]	*e*
polar surface area	*e*	*e*	*e*	*e*	<120 Å^2
functional groups	>0	*e*	*e*	*e*	*e*
molecular volume	>191 Å^3	na	na	na	na
Hutter index[h]	>0.0	na	na	na	na

[a] Taken from ref 29. [b] Taken from ref 30. [c] Taken from ref 4. [d] Taken from ref 49. [e] Not considered or no recommendation given. [f] According to the definition in ref 29. [g] Count of non-hydrogen atoms, only. [h] See ref 6.

tertiary nitrogens. The XlogP and likewise the number of hydrogen-bond donors and hydrogen-bond acceptors, respectively, as descriptors of their own appear at later branching points, although they are incorporated in other common druglikeness indices. In turn some of these druglike criteria are found at later branching points compared to other descriptors that account for specific single properties. This is insofar surprisingly, as the margins for the underlying descriptors found by the decision trees agree very well with those found by analysis of the respective substance databases: the preferred interval of the molar refractivity for drugs found by Ghose at al. is in the range between 40 and 130, whereas our (lower) margin is 40.[30] Similarly, they put the lower limits for the molecular weight at 160 and 230, respectively, whereby the later is matching our value. For the number of rings our margin (at least 1 ring system) agrees with the range suggested by Oprea (between 1 and 4 rings).[29] Moreover, we observed that the occurrence of 7-membered rings (for example in benzodiazepams) is a strong indicator of druglikeliness, since this ring size is underrepresented in nondrugs. Regarding the number of hydrogen-bond donors and hydrogen-bond acceptors as well as logP, no straightforward comparison is possible since these descriptors appear with different margins at several branching points, respectively. Thus they are used to perform subtle selections after other more decisive descriptors have been applied, rather than be used as a more general rule. Our deducted recommendation regarding the design criteria for druglike compounds are summarized in Table 8 along the ranges given by other studies. Comparison between the different types of descriptors showed that for some of the quantum chemically derived descriptors similar counterparts were found among the rapidly computable variables. This transpires when considering the strong correlation between molecular volume and molecular weight, the qualitative connection between the negative polar surface and the number of nitrogen and oxygen atoms, and likewise the relationship between the variance of the negative electrostatic potential and the number of functional groups. For the majority of quantum mechanically computed quantities, however, no such counterparts were found, underlining their uniqueness. Moreover, the performance regarding the prediction of druglike compounds depends more on the underlying descriptor set, rather than on the applied algorithm. Decision trees use a minimum of descriptors (see Table 3) that furthermore can be interpreted easily obtaining separating margins, whereas in support vector machines the corresponding support vectors spanning the dividing hyperplane comprise all of the available descriptors. Thus, decision trees are a valuable tool to identify relevant descriptors.

CONCLUSIONS

Decision trees and support vector machines have been applied to perform a gradual in silico screening for druglike compounds. By using rapidly computable descriptors, e.g., the molecular weight, the XlogP, the molar refractivity, and several druglikeness indices, up to 76% of all nondrugs can be sorted out in the first step. The application of computationally more demanding variables, e.g., SMARTS strings expressing substructures or quantum chemically derived quantities, did not improve the prediction accuracy to an extent that would justify their use in the first step. Thus, these descriptors were postponed to succeeding steps whereby up to 92% of the initial nondrugs were filtered out, while less than 19% of the actual drugs are lost. The margins for specific descriptors obtained from the decision trees agree very well with those used in existing criteria for druglike compounds. Similar indices were, however, found at later branching points than the recently introduced druglikeliness score that is based on the statistical distribution of atom pairs, highlighting its suitability for in silico screening. While it remains difficult to account for druglikeness by the means of single descriptors, the performance of the applied classifications methods benefits from such indices. We found that the prediction accuracy of decision trees is comparable to that of support vector machines with a linear kernel, demonstrating the particular feasibility of decision trees for in silico screening and simultaneously for the purpose of deriving guidelines for the design of druglike compounds.

ACKNOWLEDGMENT

We thank Yaxia Yuan and Luhua Lai for providing the training and test set compound package for XlogP. We also thank Yungki Park for technical assistance with the R program.

Supporting Information Available: Compounds from the training set, test set, and the set of new drugs as SMILES strings; SMARTS strings indicating reactive or unsuitable substructures and the SMARTS strings reflecting the chemical diversity that were used to perform the clustering; further descriptors available for the classification algorithms; and additional parameters required for the computation of XlogP, SlogP, and the molar refractivity of boron and silicon containing compounds. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) van de Waterbeemd, H.; Gifford, E. ADMET *In Silico* Modelling: Towards Prediction Paradise? *Nature Rev. Drug Discovery* **2003**, *2*, 192−204.
(2) Muegge, I. Selection Criteria for Drug-Like Compounds. *Med. Res. Rev.* **2003**, *23*, 302−321.
(3) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(4) Xu, J.; Stevenson, J. Drug-like Index: A New Approach To Measure Drug-Like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.

(5) Ertl, P. Cheminformatics Analysis of Organic Substituents: Identification of the Most Common Substituents, Calculation of Substituent Properties, and Automatic Identification of Drug-like Bioisosteric Groups. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 374−380.
(6) Hutter, M. C. Separating Drugs from Nondrugs: A Statistical Approach Using Atom Pair Distributions. *J. Chem. Inf. Model.* **2007**, *47*, 186−194.
(7) 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.
(8) Frimurer, T. M.; Bywater, R.; Nærum, L.; Lauritsen, L. N.; Brunak, S. Improving the Odds in Discriminating "Drug-like" from "Non Drug-like" Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1315−1324.
(9) Murcia-Soler, M.; Pérez-Giménez, F.; García-March, F. J.; Salabert-Salvador, M. T.; Díaz-Villanueva, W.; Castro-Bleda, M. J. Drugs and Nondrugs: An Effective Discrimination with Topological Methods and Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1688−1702.
(10) Givehchi, A.; Schneider, G. Impact of Descriptor Vector Scaling on the Classification of Drugs and Nondrugs with Artificial Neural Networks. *J. Mol. Model.* **2004**, *10*, 204−211.
(11) Wagener, M.; van Geerestein, V. J. Potential Drugs and Nondrugs: Prediction and Identification of Important Structural Features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280−292.
(12) Gálvez, J.; de Julián-Ortiz, J. V.; García-Domenech, R. General Topological Patterns of known Drugs. *J. Mol. Graphics Modell.* **2001**, *20*, 84−94.
(13) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432−2437.
(14) Brüstle, M.; Beck, B.; Schindler, T.; King, W.; Mitchell, T.; Clark, T. Descriptors, Physical Properties, and Drug-Likeness. *J. Med. Chem.* **2002**, *45*, 3345−3355.
(15) An, A.; Wang, Y. Comparison of Classification Methods for Screening Potential Compounds. Proceedings of the IEEE International Conference on Data Mining (ICDM.01), San Jose, CA, 2001; pp 11−18.
(16) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model.* **2007**, *47*, 110−114.
(17) Li, Q.; Bender, A.; Pei, J.; Lai, L. A Large Set and a Probablistic Kernel-Based Classifer Significantly Improve Druglikeness Classification. *J. Chem. Inf. Model.* **2007**, *47*, 1176−1786.
(18) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.
(19) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048−2056.
(20) Müller, K.-R.; Rätsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying "Drug-likeness" with Kernel-Based Learning Methods. *J. Chem. Inf. Model.* **2005**, *45*, 249−253.
(21) *The Merck Index*, 13th ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 2001.
(22) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(23) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.
(24) Gillet, V. J.; Willet, P.; Bradsjaw, J. Identfication of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(25) Sheridan, R. P. Finding Multiactivity Substructures by Mining Databases of Drug-Like Compounds. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1037−1050.
(26) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.
(27) Gedeck, P.; Willet, P. Visual and Computational Analysis of Structure-Activity Relationships in High-Throughput Screening Data. *Curr. Opin. Chem. Biol.* **2001**, *5*, 389−395.
(28) 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.
(29) Oprea, T. I. Property Distribution of Drug-related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(30) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.
(31) Hann, M.; Hudson, B.; Lewell, X.; Lifely, R.; Miller, L.; Ramsden, N. Strategic Pooling of Compounds for High-Throughput Screening. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 897−902.
(32) Flower, D. R. DISSIM: A Program for the Analysis of Chemical Diversity. *J. Mol. Graphics Modell.* **1998**, *16*, 239−253.
(33) Wang, J.; Lai, L.; Tang, Y. Structural Features for Toxic Chemicals for Specific Toxicity. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1173−1189.
(34) Andres, C.; Hutter, M. C. CNS Permeability of Drugs Predicted by a Decision Tree. *QSAR Comb. Sci.* **2006**, *25*, 205−309.
(35) Gepp, M. M.; Hutter, M. C. Determination of hERG Channel Blcokers Using a Decision Tree. *Bioorg. Med. Chem.* **2006**, *14*, 5325−5332.
(36) Milne, G. W. A. *Drugs: Synonyms and Properties*; 2nd ed.; Asgate: Aldershot, Hampshire, England, 2000.
(37) Janssen Pharmaceutica. *Janssen Chimica 88−90*; Janssen Pharmaceutica: Brüggen, Germany, 1988.
(38) Comparative Evaluation of Prediction Algorithms. http://www.coepra.org (accessed Nov 23, 2007).
(39) Banck, M.; Bresciani, F.; Bréfort, J.; Clark, A.; Corkery, J.; Favre-Nicolin, V.; Fontaine, F.; Gillies, M.; Gillilan, R.; Goldman, B.; Hassinen, T.; Herger, B.; Hutchison, G.; Kebekus, S.; Kruus, E.; Leitl, E.; Mathog, D.; Morley, C.; Murray-Rust, P.; Nicholls, A.; Patchkovskii, S.; Reith, S.; Richard, L.; Sayle, R.; Shah, A.; Stahl, M.; Tolbert, B.; Walters, P.; Wolinski, P.; Wegner, J. *Open Babel, version 1.100.2*. http://openbabel.sourceforge.net (accessed Sep 5, 2005).
(40) Wang, R.; Gao, Y.; Lai, L. Calculating Partition Coefficient by Atom-Additive Method. *Perspect. Drug Discovery Des.* **2000**, *19*, 47−66.
(41) Viswanadhan, V. N.; Ghose, A. K.; Rebankar, 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 Occuring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(42) Willet, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(43) Rishton, G. M. Reactive Compounds and in vitro False Positives in HTS. *Drug Discovery Today* **1997**, *2*, 384−386.
(44) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular-Model: 76. AM1-A new General-Purpose Quantum-Mechanical Molecular-Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(45) Rauhut, G.; Alex, A.; Chandrasekhar, J.; Steinke, T.; Sauer, W.; Beck, B.; Hutter, M.; Gedeck, P.; Clark, T. *VAMP, version 6.5*; Oxford Molecular: Erlangen, Germany, 1997.
(46) Baker, J. An Algorithm for the Localization of Transition-States. *J. Comput. Chem.* **1986**, *7*, 385−395.
(47) Dimitriadou, E.; Kurt Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, W. *e1071, version 1.5-13*. http://cran.r-project.org (accessed Jan 7, 2006).
(48) Bates, D.; Chambers, J.; Dalgaard, P.; Falcon, S.; Gentleman, R.; Hornik, K.; Iacus, S.; Ihaka, R.; Leisch, F.; Lumley, T.; Maechler, M.; Murdoch, D.; Murrell, P.; Plummer, M.; Ripley, B.; Sarkar, D.; Temple Lang, D.; Tierney, L.; Urbanek, S. R. *version 2.3.1*. http://cran.r-project.org (accessed Jan 7, 2006).
(49) Walters, W. P.; Murcko, M. A. Library Filtering Systems and Prediction of Drug-Like Properties. *Meth. Principles Med. Chem.* **2000**, *10*, 15−30.
(50) Kelder, J.; Grootenhuis, P. D. J.; Bayada, D. M.; Delbressine, L. P. C.; Ploemen, J.-P. Polar Molecular Surface as a Dominating Determinant for Oral Absorption and Brain Penetration of Drugs. *Pharm. Res.* **1999**, *16*, 1514−5119.
(51) Zheng, S.; Luo, X.; Chen, G.; Zhu, W.; Shen, J.; Chen, K.; Jiang, H. A New Rapid and Effective Chemistry Space Filter in Recognizing a Druglike Database. *J. Chem. Inf. Model.* **2005**, *45*, 856−862.
(52) Oprea, T. I.; Allu, T. K.; Fara, D. C.; Rad, R. F.; Ostopovici, L.; Bologa, C. G. Lead-like, Drug-like or "Pub-like": How Different are They? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 113−119.
(53) Karakoc, E.; Shinalp, S. C.; Cherkasov, A. Comparative QSAR- and Fragments Distribution Analysis of Drugs, Druglikes, Metabolic Substances, and Antimicrobial Compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2167−2182.
(54) Lipinski, C. A. Drug-like Properties and the Causes of poor Solublity and poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235−249.
(55) Meyer, A. Y. The Size of Molecules. *Chem. Soc. Rev.* **1986**, *15*, 449−475.
(56) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. Using Theoretical Descriptors in Quantitative Structure-Property Relation-

ships: Some Distribution Equilibria. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293−1301.
(57) Breindl, A.; Beck, B.; Clark, T.; Glen, R. C. Prediction of the n-Octanol/Water Partition Coefficient, logP, Using a Combination of Semiempirical MO-Calculations and a Neuronal Network. *J. Mol. Model.* **1997**, *3*, 142−155.
(58) Beck, B.; Glen, R. C.; Clark, T. VESPA: A New, Fast Approach to Electrostatic Potential Derived Atomic Charges from Semiempirical Methods. *J. Comput. Chem.* **1997**, *18*, 744−756.
(59) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular 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.
(60) Kier, L. B.; Hall, L. H. *Molecular Structure Description: The Electrotopological State*; Academic Press: San Diego, CA, 1999.
(61) Katritzky, A. R.; Mu, L.; Lobanov, V. S.; Karelson, M. Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics. *J. Phys. Chem.* **1996**, *100*, 10400−10407.
(62) Osmialowski, K.; Halkiewicz, J.; Kaliszan, R. Quantum Chemical Parameters in Correlation Analysis of Gas-Liquid Chromatographic Retention Indices of Amines. 2. Topological Electronic Index. *J. Chromatogr., A* **1986**, *361*, 63−69.
(63) Meyer, Y. A. Molecular Mechanics and Molecular Shape. V. On the Computation of the Bare Surface Area of Molecules. *J. Comput. Chem.* **1988**, *9*, 18−24.
(64) Jorgensen, W. L.; Duffy, E. M. Predicting Drug Solubility from Monte Carlo Simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.

CI700351Y