Separating Drugs from Nondrugs: A Statistical Approach Using Atom Pair Distributions

Michael C. Hutter*

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

Received July 31, 2006

A computational approach to quantify the druglike character of chemical compounds is presented. For this purpose, the distribution of atom types and their pair-wise combinations in known drugs and nondrugs was examined. Statistical analysis of the occurrence probabilities was used to derive a drug-likeliness score on a logarithmic scale. "Typical" pharmaceutical agents exhibit scores greater than 0.3, while for ordinary substances, values below 0 are expected. Although any kind of fitting or error minimization scheme is absent in this method, confirmed drugs are predicted with an accuracy of at least 71%. Many falsely predicted nondrugs were found to closely resemble actual drugs or to contain unsuitable substitution patterns that can easily be ruled out by applying medicinal knowledge. As the outlined method is computationally inexpensive, this drug-likeliness score can therefore be used as a filter for the in silico screening of large substance databases.

1. INTRODUCTION

The computer-assisted design of new potential drugs requires effective means and ways to filter out unsuitable compounds as early as possible. Because more than 80% of all failures of commercial drugs can be attributed to inappropriate absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties despite in vitro and in vivo testing, in silico screening methods more and more become indispensable tools during preclinical development.[1] Beyond the prediction of solubility, absorption, distribution, metabolism, elimination, and toxicity separately by individual methods, the challenge remains to predict the druglikeness of a substance per se. In the ideal case, such approaches should be able to distinguish potential drugs from nondrugs in one step rather than applying successive filtering steps for each the ADMET issues. This process is of particular interest upon setting up virtual libraries for further use, for example, high-throughput screening. The basic question behind such a classification scheme is, which properties make pharmaceutical drugs different from ordinary chemicals? An analysis of large data collections such as the World Drug Index yielded guidelines for the design of orally administered drugs, known as Lipinski's rule of five.[2] The occurrence of common structural features such as frameworks and side chains was investigated by Bemis and Murcko.[3,4] Other knowledge-based approaches that make use of the distribution of features in substance databases were reported by Wang and Ramnarayan, Ghose et al., and Xu and Stevenson, as well as by Oprea.[5−8] On the basis of the occurrence of certain atom types in compounds, Sadowski and Kubinyi trained a neural network to generate a scoring scheme that was able to separate drugs and nondrugs.[9] Related approaches using various descriptors were applied by Givehchi and Schneider, as well as by Clark and co-workers also including unsupervised self-organizing maps, and furthermore by Murcia-Soler et al., Frimurer et al., and Ajay et al.[10−14] A linear discriminant analysis based solely on topological indices was performed by Gálvez et al.[15] Further computational methods applied to classify compounds include machine learning algorithms such as support vector machines and decision trees.[16−19] The latter method is of particular use for medicinal chemists because guidelines for the design of compounds with desired properties can easily be extracted from a decision tree, for example, the presence and numerical count of certain chemical groups or substructures. Furthermore, the occurrence of unsuitable molecular fragments can be used to filter out problematic compounds beforehand, for example, by using SMARTS.[20] Collections of such reactive, toxic, or hard to synthesize substructures are given by Flower and by Hann et al.[21,22] Conversely, a combination of functional groups was used to generate so-called pharmacophoric points as a filter for druglikeness.[23] It was observed that nondrugs contain fewer functional groups than druglike compounds.

All of the quoted studies exploit the (assumed) unequal distribution of certain features between drugs and nondrugs. The approach outlined in this work is based on the assumption that certain atom pair combinations occur with a different frequency in druglike molecules compared to other substances. This hypotheses is tested by statistical means. For this purpose, some aspects of the well-established methods for determining amino acid sequence alignments from similarity matrices are extended to derive a drug-likeliness score for molecules. Atom pairs were used earlier to describe the similarity between molecules or to correlate them with biological activity.[24] This atom pair concept was extended furthermore to topological torsions where four consecutively bonded non-hydrogen atoms are taken into account.[25] A related approach is realized in the PASS program that applies

* Author phone: ++49 +681 302 64178; fax: ++49 +681 302 64180; e-mail: michael.hutter@bioinformatik.uni-saarland.de.

10.1021/ci600329u CCC: $37.00 © 2007 American Chemical Society
Published on Web 12/02/2006

MNA descriptors that account for the multilevel neighborhood of atoms also in the context of drug/nondrug discrimination.[26] The vicinity of an atom is also used in the multilevel chemical compatibility approach of Wang and Ramnarayan that comprises seven different atom types for all nonhydrogen atoms.[5] The mentioned approaches apply atom types that should reflect intrinsic properties associated with the occurrence of a specific atom. It is easy to see that the simple element number alone cannot fulfill such requirements that go along with fragment-based approaches for the estimation of logP or the molar refractivity, for example.[27] To describe all relevant chemical groups appropriately, here, atom types are applied that are also used in a force field.

2. METHOD

2.1. Compound Data Sets. The chemical substances used in this study were collected from the Merck Index,[28] G. Milne's compilation of drugs,[29] and a commercial vendor catalog of chemicals.[30] Only those compounds were taken into account that contain elements from the group of C, N, O, F, Cl, Br, I, S, P, Si, and B. To ensure a net neutral charge, carboxylates were converted into their neutral counterparts by replacing Na$^+$ and K$^+$ with hydrogen. Conversely, the presence of a negative counterion in the vicinity of quaternary nitrogens is characterized by "attaching" Cl$^-$ to this atom. Inner salts were kept unchanged. Mixtures of two or more compounds were not considered in order to ensure that only single molecules are present. Likewise, HCl was removed from the entries where necessary.

The compounds were separated into drugs and nondrugs applying the following criteria:

1. Nondrugs must not be assigned to a pharmaceutical category but can be a diagnostic aid, pharmaceutical aid, dye, pigment, solvent, surfactant, insect repellent, sun screen, ultraviolet screen, sweetener, or flavoring. Usually, reactive or toxic compounds are removed from the training and test sets.[9] Here, however, such compounds are included in the nondrug group to capture the features of those substances.

2. Drugs are characterized by affiliation to a typical pharmaceutical category, but excluding intravenous- and inhalative-administered compounds as well as those that exhibit a toxicological function such as pesticides, insecticides, rodenticides, miticides, plant growth inhibitors, and acaricides.

Applying these selection criteria yielded a total of 2713 pharmaceutical drugs and 1373 nondrugs. These were further partitioned into nonoverlapping training and test sets whereby the respective test sets comprised about 8% of all compounds. The test set for the nondrug group was based on the compilations of nondrugs by Murcia-Soler et al. and by Gálvez et al. after a critical re-evaluation of possible pharmaceutical indications.[14,15] Compounds for the test set of the drug group were chosen in a way that reflects the proportional distribution of all agents among the therapeutic categories. Because of the absence of conventional descriptors that would allow a cluster-based splitting of the compounds into training and test sets, this step was performed manually in order to represent occurring structural motives in the sets adequately. The statistical theory as outlined below is furthermore designed to handle "asymmetries" arising from unequal numbers of molecules in the two compound classes. Each occurring feature is normalized, and thus, its frequency is independent from the number of considered molecules. This is a conceptual advantage compared to other partitioning schemes that allows the handling of otherwise imbalanced data sets.

Table 1. Explanation of the Atom Types Used

type	description	type	description
C4	Csp3	O2	oxygen in C−O−H and C−O−C
C3	Csp2 alkene	O1	carbonyl oxygen =O
CO	Csp2 carbonyl	OF	Osp2 furan
C2	Csp2 alkyne and C=C=O	OC	carboxylate oxygen
CP	Csp3 cyclopropane	OE	epoxy oxygen
C+	carbonium	ON	amine oxide oxygen
CZ	Csp2 cyclopropene	O?	any other oxygen
CA	Csp2 aromatic	B3	trigonal boron
CB	Csp3 cyclobutane	B4	tetragonal boron
CC	Csp2 cyclobutene	Si	silicon
CD	Csp2 cyclobutanone	P	phosphine phosphorus >P−
CE	Csp2 cyclopropanone	P5	pentavalent phosphorus
C?	any other carbon	P?	any other phosphorus
N3	Nsp3	S2	sulfide −S−
N2	Nsp2 amide	S+	sulfonium >S+
N1	Nsp	SO	sulfoxide >S=O
NA	nitrogen in =N−, azo, pyridine	S4	sulfone >SO2
NH	Nsp3 ammonium	SA	Ssp2 thiophene
NP	Nsp2 pyrrole	S?	any other sulfur
NB	nitrogen in −N=N−O, azoxy	F	fluorine
NZ	azide central nitrogen	Cl	chlorine
NO	NO2 nitrogen	Br	bromine
NC	nitrogen in =N−,imine, oxime	I	iodine
N?	any other nitrogen		

2.2. Theory. The basic assumption behind the following approach made here is that drugs and nondrugs exhibit a different distribution of certain features. In principle, these features can be of any kind or property that can be quantified numerically, such as topological descriptors, logP, count of elements, and so on. For example, Lipinski's rule of five uses the count of hydrogen-bond donors and acceptors to classify orally administered drugs.[2] Here, the molecular features are represented by the occurrence of atom type combinations. Sadowski and Kubinyi used solely the occurrence of 92 atom types as input for a neural net to generate a scoring scheme.[9] These atom types stem from a fragment-based approach for the prediction of logP and thus reflect the necessary amount of types to describe all relevant chemical groups. Such an exhaustive description of atom types is also required in molecular force fields, and thus, corresponding atom types should be suited for the description of substances. Therefore, the atom types of the MM+ force field as implemented in HYPERCHEM were used here.[31] For carbon, 13 different atom types are used, 11 for nitrogen, 7 for oxygen, 6 for sulfur, 3 for phosphorus, 2 for boron, 1 for silicon, and 1 for each of the halogen atoms. These include default atom types for carbon, nitrogen, oxygen, phosphorus, and sulfur to account for circumstances where no other matching atom type can be assigned. All atom types are given in Table 1.

Hydrogen was excluded for two reasons: first, hydrogen atoms occur much more frequently than other elements, and second, the chemical properties of a hydrogen atom depend on its vicinity. For example, the chemical nature of the substituent R of a carboxylic group R−COOH determines the according pK_a quite individually. Thus, it is unnecessary to include atom types for hydrogen in this approach.

The assumption that the distribution of certain atom types (and their pair-wise combinations) is different between drugs and nondrugs has to be verified. For this purpose, the respective likelihood is calculated. This is the probability p_i of an event i happening under one hypothesis divided by the probability of the same event happening under an alternative hypothesis. Here, it is assumed that the atom types are arbitrarily distributed in contrast to the alternative hypothesis that the distribution is purely random.

The frequency p_i of finding the atom type i is thus the sum of all concurrencies of atom type i in all N considered molecules divided by the sum of all occurrences of all n different atom types in all N molecules. This takes into account that there are presumably more carbon atoms than nitrogen atoms present, for example.

$$p_i = \frac{\sum^{N} i}{\sum^{N}\sum_{i=1}^{n} \text{atom types}} \tag{1}$$

Likewise, the frequency q_{ij} of finding a pair of atoms of types i and j is the sum of all occurrences of the respective ij pair in all N molecules divided by the sum of the occurrences of all possible pairing combinations of atom types i and j in all N molecules.

$$q_{ji} = \frac{\sum^{N} ij}{\sum^{N}\sum_{i=1}^{n}\sum_{j=i}^{n} \text{atom pair } ij} \tag{2}$$

It is easy to see that summation of all individual frequencies p_i must yield 1, as this is the total probability. Likewise, the summation of all frequencies q_{ij} over all molecules yields 1.

$$\sum_{i}^{N}\sum_{j}^{N} q_{ij} = 1 \tag{3}$$

To test if an atom pair ij is nonrandomly distributed among the considered molecules, we have to compare its probability (expressed by q_{ij}) to the probability of finding this atom pair by chance. The latter event is the so-called *null hypothesis* that is given as the product of the individual probabilities p_i and p_j of the atom types i and j. The comparison is done by dividing q_{ij} by p_i times p_j, giving rise to a relative probability S_{ij}'.

$$S_{ij}' = \frac{q_{ij}}{p_i p_j} \tag{4}$$

Because S_{ij}' may cover a wide numerical range from 0 to infinity depending on the individual frequencies, its logarithm is used instead and is called a *log odds* score.

$$S_{ij} = \ln \frac{q_{ij}}{p_i p_j} \tag{5}$$

Now, values of S_{ij} greater than 0 indicate a pairing frequency q_{ij} that is larger than expected by chance, while negative values mean lower frequencies than random. Rearranging eq 5 while making use of eq 3 we therefore get

$$\sum_{i}^{N}\sum_{j}^{N} q_{ij} = 1 = \sum_{i}^{N}\sum_{j}^{N} p_i p_j \, \mathrm{e}^{S_{ij}} \tag{6}$$

Because the individual frequencies p_i and p_j sum up to 1, respectively (see above), eq 6 simplifies to

$$\sum_{i}^{N}\sum_{j}^{N} \mathrm{e}^{S_{ij}} = 1 \tag{7}$$

as a consistency condition that has to be fulfilled to account for the following reasons: First, atom pair combinations that do not occur or are very rarely found among the considered compounds lead to some problems. According to eq 5, their corresponding S_{ij} elements get large negative values. An even more severe problem is encountered if a certain atom type i is not found, giving rise to $p_i = 0$. To circumvent the subsequent division by zero in eqs 4 and 5, the initial elements S_{ij}' are assigned a minimum probability of 1×10^{-4}. Normalization of S_{ij} to fulfill eq 7 is achieved by introducing a scaling factor λ in eq 6:

$$\sum_{i}^{N}\sum_{j}^{N} q_{ij} = 1 = \sum_{i}^{N}\sum_{j}^{N} p_i p_j e^{\lambda S_{ij}} \tag{8}$$

Because λ cannot be determined analytically, a numerical minimization using the iterative Newton method is performed:

$$\lambda_{n+1} = \lambda_n - \frac{f(\lambda_n)}{f'(\lambda_n)} \tag{9}$$

where

$$f(\lambda) = \sum_{i}^{N}\sum_{j}^{N} p_i p_j \, \mathrm{e}^{\lambda S_{ij}} - 1 = 0 \tag{10}$$

This iteration converges rapidly if an initial value of 1 is chosen for λ. Elements of the matrix that contain the assigned minimum probability of $\ln(1 \times 10^{-4})$ are excluded from the normalization because, otherwise, the random probabilities are shifted away from 0. All other matrix elements S_{ij} are multiplied by λ. Values for λ were found to be in the range of 1.013−1.225 for all of the considered matrices. This is the second reason for this normalization procedure because matrices can by derived from groups containing different numbers of compounds. For example, there are many more nondrugs than actual drugs present today. In contrast to conventional regression analysis, it is thus possible to handle "asymmetric" data sets as the occurrence of a certain feature is throughout considered as its respective probability (see eqs 1 and 2) and thus is independent from the number of compounds.

So far, S_{ij} are elements of a matrix that closely resembles the approach of Karlin and Altschul for the statistical analysis of sequence features and amino acid exchange matrices.[32] As we are dealing with single molecules instead of pair-wise alignments of amino acid sequences, the appropriate scoring function has to be different.

The likeliness score L' for a compound is calculated by summing up the corresponding matrix elements S_{ij} for all atom pairs of i and j present in that molecule, divided by the number of atom pairs M in the molecule. Division by M is required to account for the higher number of atom pairs in larger molecules. In contrast, two sequence alignments are equal in length.

$$L' = \frac{1}{M}\sum_{i=1}^{n}\sum_{j=1}^{n}\begin{cases} S_{ij} \text{ if } ij \text{ in molecule} \\ 0 \text{ else}\end{cases} \tag{11}$$

Positive values of L' would indicate a higher probability of that compound being similar to those compounds used to derive the matrix **S** than by chance.

To distinguish between two classes A and B (i.e., drugs and nondrugs), the respective matrices $\mathbf{S}^{\mathbf{A}}$ and $\mathbf{S}^{\mathbf{B}}$ have to be generated using a (large) set of molecules from each class following the procedure outlined above. We would then be able to determine the two different likeliness scores on the basis of the two individual matrices. This leaves us with the task of evaluating two numerical scores rather than one. Alternatively, we can, however, generate a difference matrix **D** by subtracting the corresponding matrix elements from each other.

$$D_{ij} = S_{ij}^{\mathrm{A}} - S_{ij}^{\mathrm{B}} \tag{12}$$

This is allowed if the condition given in eq 7 is fulfilled for both matrices and also ensures that random probabilities remain 0. This is achieved by applying individually determined values of λ for both matrices in eq 8. Now, positive matrix elements of **D** signal atom pair combinations that are over-represented in class A compared to those in class B. The calculated likeliness score according to eq 11 using the difference matrix can be evaluated rather easily: Positive values indicate a tendency for class A, whereas negative values suggest affiliation to class B.

Up to now, we have considered occurrence frequencies of atom types and covalently bonded atom pairs (1−2 interactions) only. The concept of atom pair frequencies can, however, easily be extended to bond angles (1−3 interactions), torsions (1−4 interactions), and higher interactions. Here, atom pair combinations are evaluated up to 1−6 interactions. For each of these atom pair interactions, a separate difference matrix D_{ij}^{k} is computed. Adding up the corresponding score from each of these matrices yields the total likeliness score L.

$$L = \sum_{k=1}^{6} L' = \frac{1}{M}\sum_{k=1}^{6}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\begin{array}{l} D_{ij}^{k} \text{ if } ij \text{ in molecule} \\ 0 \text{ else}\end{array}\right\}\right) \tag{13}$$

This means that for each of the two compound classes the six interaction matrices S_{ij} are determined first using the respective training sets. Subsequently, the six difference matrices D_{ij}^{k} between corresponding interaction matrices from the two classes are calculated. The likeliness score L for a single compound can now be calculated from eq 13 by adding up the individual matrix entries for occurring atom pair combinations.

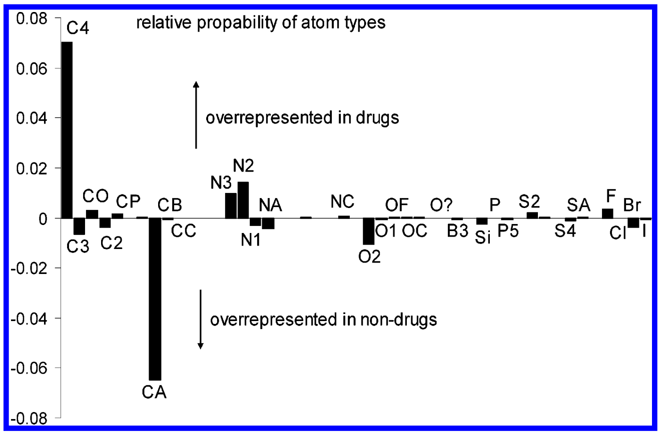

Figure 1. Difference of the relative probabilities of finding a certain atom type in drugs compared to that in nondrugs (1−1 interaction). Only those atom types are marked that exhibit a difference of 1 × 10^{-4} or more. The full list of all atom types is given in Table 1. Positive values indicate over-representation in drugs, while negative values denote over-representation in nondrugs.

Because each matrix has a dimension of n times n (with $n = 47$ being the number of atom types), the apparent number of parameters seems to be rather large at first glance. A considerable amount of the matrix elements is, however, 0 as many atom pair combinations were not observed in the compounds and therefore do not enter the calculation of the score. Furthermore, none of the matrix elements are subject to any kind of fitting procedure as in conventional methods but solely reflect the statistical distribution of the observed atom pair distribution. Therefore, the usual concept of descriptor space and degrees of freedom cannot be applied to the present approach.

2.3. Timing. Generation of the matrices and computation of the respective likeliness scores for the total of 4086 compounds using PERL scripts took less than 2 min on one Intel Pentium IV processor running at 2.8 GHz.

3. RESULTS AND DISCUSSION

To test the assumption that certain atom types are distributed differently in drugs and nondrugs, their relative probability was calculated according to eq 1. Subtraction of the respective probabilities for the nondrug group from those of the drug group yields the likeliness score for the atom types (shown in Figure 1). Surprisingly, the largest differences occur for carbon atoms, namely the sp^3 carbon (C4) and the aromatic sp^2 carbon (CA). This means that aromatic rings (e.g., benzene and naphthalene) are markedly overrepresented in nondrugs despite being the most common ring fragments in drugs.[3] Conversely, fragments composed of multiple CH_2 groups (that contain the sp^3-hybridized carbons of type C4) seem to be a typical feature of druglike compounds. These can be either linkers or rings, such as cyclohexane. Differences for other elements are less emphasized. For nitrogen, a preference of sp^3 nitrogens (N3) and amide nitrogens (N2) in drugs is found compared to sp nitrogens (N1), for example, in azido and nitrile substituents, and aromatic nitrogens (NA) in nondrugs. Likewise alcoholic, ether, and ester oxygens (O2) prevail in nondrugs. As expected from the small number of silicon-containing pharmaceutical agents, this atom type is found almost exclusively in nondrugs. The same situation also applies to boron. The trend for the halogens is mixed; especially, chlorine is equally distributed, while slight preferences are seen for fluorine and bromine.

Analysis of the pair-wise combinations of these atom types is the next step. Shown in Figure 2a are the interactions that are considered. While the 1−1 interaction corresponds to the above-mentioned distribution of atom types, the easiest and most obvious is the 1−2 interaction that corresponds to a pair of covalently bonded atoms. Computing the logarithmic probability S_{ij} of finding an atom of type i covalently bonded to an atom of type j among a group of compounds is done according to eq 4. These probabilities are elements of a symmetrical matrix which is subsequently normalized. Subtracting the respective matrix elements of the nondrug matrix from those of the drug matrix (eq 12) yields the difference matrix. The numerical range of the according elements is shown color coded in Figure 2b. Similar to Figure 1, positive values indicate an over-representation of the particular atom pair combination in drugs whereas negative values denote over-representation in nondrugs. For example, the strong negative value of −9.157 for the 1−2 interaction of CO and Cl stems from reactive carbonyl chlorides that are exclusively present in nondrugs. Conversely, cyclopropane rings bonded to aromatic rings give rise to a value of 7.043 for the CP−CA interaction. This structural motif is predominately found in drugs.

Extension of this concept of pair-wise interactions of atom types to 1−3, 1−4, 1−5, and 1−6 interactions yields the according difference matrices shown in Figure 2c−f. Comparison of these matrices with respect to each other reveals subtle differences regarding the distribution of atom pair combinations between drugs and nondrugs. Particularly, the 1−6 interaction matrix exhibits the most differences of all matrices, solely by visual inspection. A more precise numerical estimate is achieved by summing up the absolute values of all matrix elements in the respective matrix. This confirms the sequence 1−6 > 1−4 > 1−5 > 1−3 > 1−2. One would expect that matrices of higher interaction should always show larger differences than those of the 1−2 interaction, simply by the number of combinatorial possibilities. The occurrence of cyclic structures, for example, benzene rings where carbon has only three substituents, however, diminishes the possible combinations. Furthermore, the most common atom in ring systems is either an aliphatic or an aromatic carbon. This explains why the matrix of the 1−5 interactions shows a smaller difference than the 1−4 matrix.

As outlined in the Method section, the matrix elements can be used to compute a likeliness score for a molecule that expresses the logarithmic probability of that molecule belonging to the class from which the respective matrix was derived. Using the above-mentioned difference matrices results in a drug-likeliness score. Here, positive values indicate a druglike agent whereas negative values denote attributes of nondruglike molecules. Upon calculating the likeliness scores for each interaction, it turned out that only the respective sum of all scores according to eq 13 yielded useful results. The distribution of the individual scores (1−1 to 1−6) and the total drug-likeliness score among the range of compounds is shown in Figure 3 for nondrugs and drugs. Also shown is the coverage of all compounds (in percent) up to the respective drug-likeliness score. For both drugs and nondrugs, peaks on both sides of the separating margin at 0.0 are visible for all scores based on the individual interactions. The total drug-likeliness score (in the front), however, shows less-emphasized peaks and a slower descent of its tails. Particularly for drugs, the tail descends much faster to zero in the (wrong) negative region than conversely in the positive region of the nondrug group. This is also reflected by the degree of coverage of the compounds. Summing up all nondrugs with negative scores yields only 57.2% at 0.0. Including wrongly predicted molecules with positive scores yields 67.2% at 0.1 and 70.7% at 0.2. The situation looks better for drugs. Here, all correctly predicted agents with drug-likeliness scores greater than 0 cover already 77.4% of all compounds in this group. At −0.1, the coverage has increased to 80.1% and further to 84.3% at −0.2. A detailed list of the computed drug-likeliness scores for the individual interactions (1−1 to 1−6) for the 328 compounds of the test sets is provided as supplementary information. The obtained prediction accuracies for separating drugs from nondrugs using the various likeliness scores are summarized in Table 2 for all four sets of compounds. A score of 0.0 was used throughout as a separating margin. From Table 2, it is easily visible that only the drug-likeliness score is practicable for separating drugs from nondrugs. The two other likeliness scores that are solely based on the drugs and nondrugs matrices do not provide meaningful selection criteria on their own as their quality in prediction breaks down for compounds of the respective other class. Obviously, only the difference of the probabilities of the atom pair combinations between drugs and nondrugs provides significant information. Except for the nondrugs test set, corresponding χ^2 tests yielded significance levels of $p < 0.001$ or better for the nondrug training set, the drug test set, and the drug training set. A further cross-validation using the difference matrices where the matrix elements ij were replaced by ji in each interaction matrix yielded similar values (drugs test set 6.0%, drugs training set 9.9%, nondrugs test set 10.0%, and nondrugs training set 26.0%) to those when applying the matrices of the respective other class.

It should be emphasized at this point that the approach applied here is solely derived from the statistical distribution of atom types and their pair-wise combinations present in the compounds and does not contain any fitting or error minimization scheme. Thus, there is a statistical difference in the distribution of these features between drugs and nondrugs which proves the assumption made at the beginning of the section.

Of interest are the falsely predicted compounds. Examples of molecules that exhibit high positive values in the group of nondrugs are shown in Figure 4. Typical failures are encountered for structural patterns that are rarely found in nondrugs (e.g., the five-membered ring of parabanic acid) or occur frequently in drugs. For example, the six-membered ring of barbituric acid is typical for sedatives. High drug-likeliness scores are also observed for hexachlorobenzene and a number of very small compounds such as acetaldioxime. The extreme halogen substitution of hexachlorobenzene is again rarely found in nondrugs as insecticides and pesticides were not taken into account. Such unwanted molecules can, however, easily be detected applying empirical selection criteria, for example, by the use of SMARTS.[21,22] Interestingly, reactive agents that were present in the

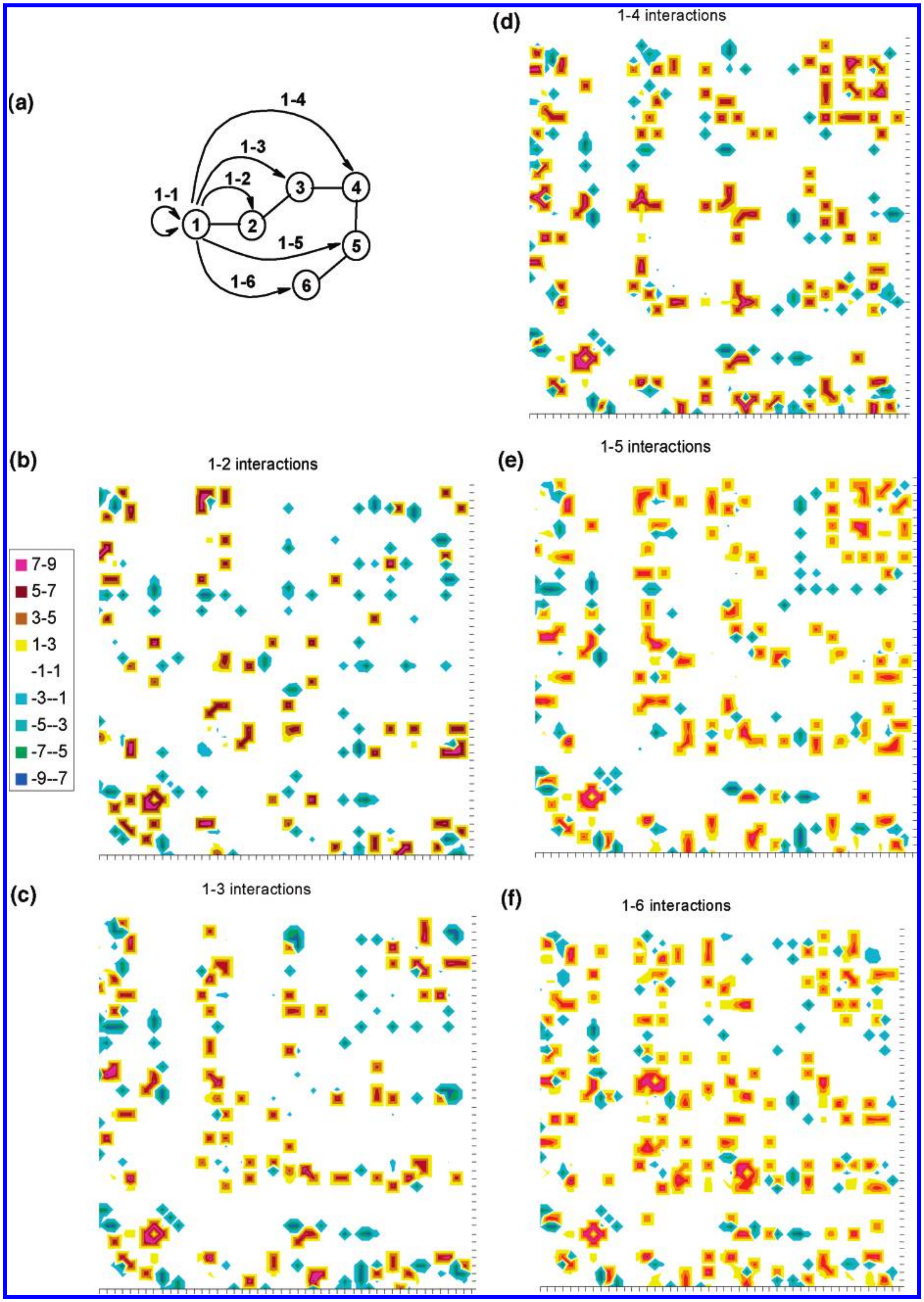

Figure 2. (a) Definition of the atom pair combinations used. The 1−1 interactions refer to the atom-type probabilities. (b−f) 1−2 interactions refer to two covalently bonded atoms, 1−3 interactions to bond angles, 1−4 interactions to torsions, and higher interactions to those between atoms that are in a continuous sequence of covalently bonded atoms. The elements of the difference matrices for these atom pair combinations are represented graphically. The logarithmic probability of finding a certain pair of two atom types is color coded: positive values indicate combinations that are over-represented in drugs (yellow to purple); negative values indicate those that are over-represented in nondrugs (cyan to dark blue), whereas white spaces represent an approximately even distribution between drugs and nondrugs. Labeling of the axes with corresponding atom types has been omitted for clarity but refers to the *x* axis in Figure 1 (from left to right and from bottom to top: carbons, nitrogens, oxygens, boron, silicon, phosphorus, sulfurs, and the halogens).

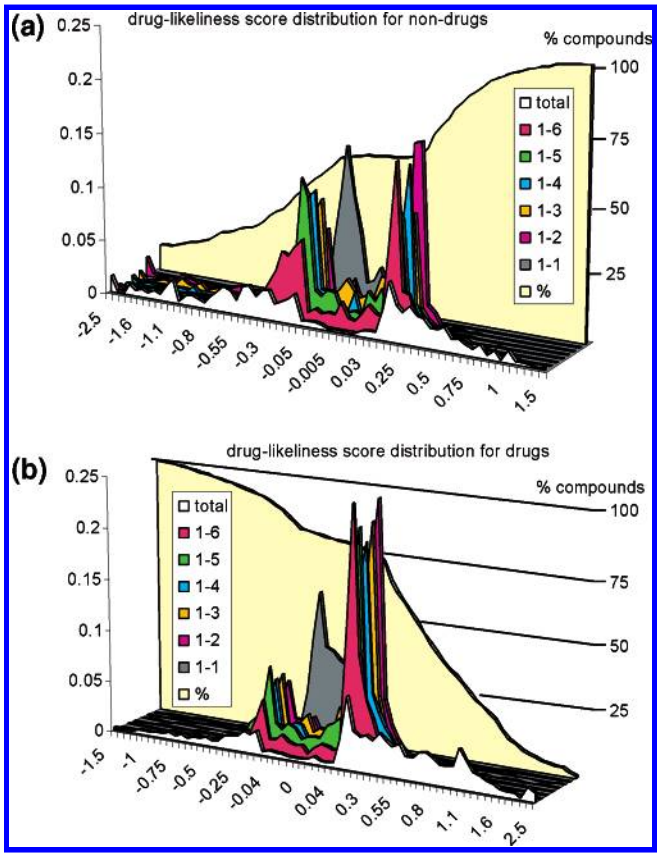

Figure 3. Distribution of the drug-likeliness score among the compounds using the various atom pair combinations: (a) for the nondrugs and (b) for the drugs. Also shown is the degree of coverage of all compounds given in percent. Please note that the resolution of the *x* axis is higher between −0.1 and +0.1 than for the remaining data range.

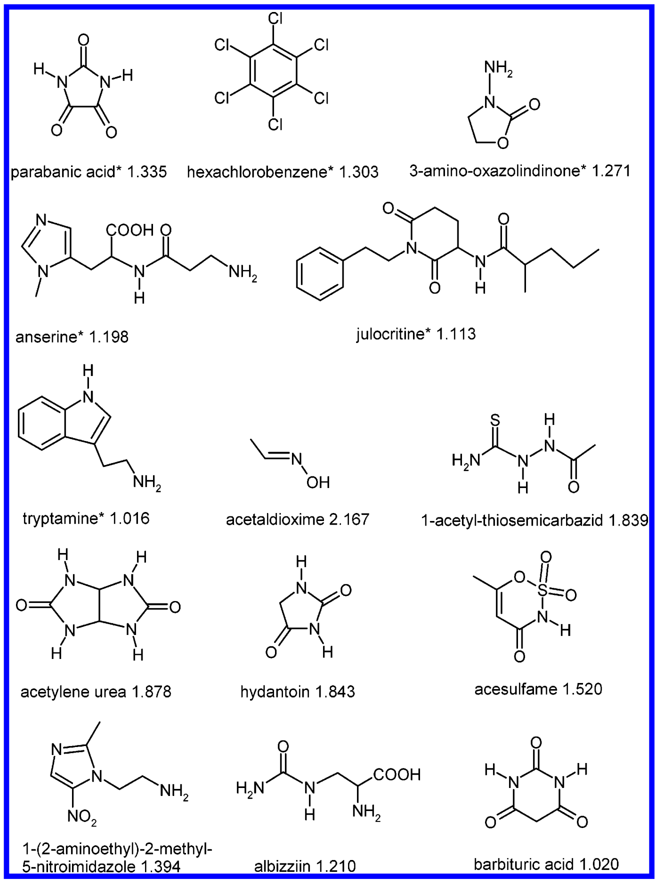

Figure 4. Examples of falsely classified nondrugs with particular high drug-likeliness scores. Large positive values indicate an increasing probability of that compound possessing druglike character. An asterisk denotes compounds that were part of the test set.

Table 2. Computed Prediction Accuracy of the Total Likeliness Scores

compound set	number of compounds	using difference matrices[a]	using nondrugs matrices[b]	using drugs matrices[b]
drugs test	218	71.1%	7.3%[c]	98.6%
drugs training	2495	77.4%	9.2%[c]	98.8%
nondrugs test	110	40.9%	87.3%	6.4%[c]
nondrugs training	1263	57.2%	90.2%	15.3%[c]

[a] Yielding the drug-likeliness score. [b] Yielding the likeliness score of being a member of that compound class. [c] These values correspond to a cross-validation scenario where the likeliness score for the compounds of one class was calculated using the matrices of the respective other class.

nondrugs training set, such as carbonyl chlorides, are successfully recognized in the test set.

Conversely, Figure 5 shows examples of drugs with strong negative drug-likeliness scores. Again, rare combinations of atom types appear: flusilazole is one of the few drugs that contains silicon. Further typical failures comprise the occurrence of rare fragments (cyclopropane rings sharing a common bond with five- or six-membered rings such as in drospirenone) and predominately hydrocarbon skeletons with few substituents (perhexilene and dymanthine). The latter characteristic is frequently found in drugs that penetrate the blood−brain barrier and/or show high logP values (gabapentin, tedisamil, cyclexedrine, and mecamylamine). As a general trend, steroids and steroidlike agents receive drug-likeliness scores below zero because of the presence of the large hydrocarbon skeleton. Their actual pharmaceutical functionality is, however, determined in a rather delicate way by a few substituents. Similar to the situation in the nondrug group, molecules with a low molecular weight (fosphosal and ethyl chloride) are difficult to determine as they contain only a few feature combinations. Thus, low drug-likeliness scores are found in general for rather “untypical” drugs. These also comprise agents that are derived from naturally occurring substances (e.g., atropine and yohimbine).

Furthermore, the test set for nondrugs used in this study contains many compounds with structures that closely resemble actual drugs (e.g., aminopyrone, barbituric acid, meteloidine, and tryptamine), which therefore explains the lower prediction accuracy (40.9%) compared to that for the drug test set (87.3%). Similar results were found by Sadowski and Kubinyi where compounds with high scores showed druglike character despite being nondrugs and vice versa.[9] Table 3 contains the calculated drug-likeliness scores for a number of top-selling drugs adopted from the same authors and complemented by some more recent agents. Apparent is again the failure for steroidlike agents, the statins, imatinib, and orlistat because of their extensive hydrocarbon framework with few substituents. The low score of salbutamol can be attributed to its low complexity. For all other drugs, scores well above 0 are obtained. Surprising is the high value for sibutramine that is due to the unusual vicinity of the cyclobutane ring. A further comparison with other studies

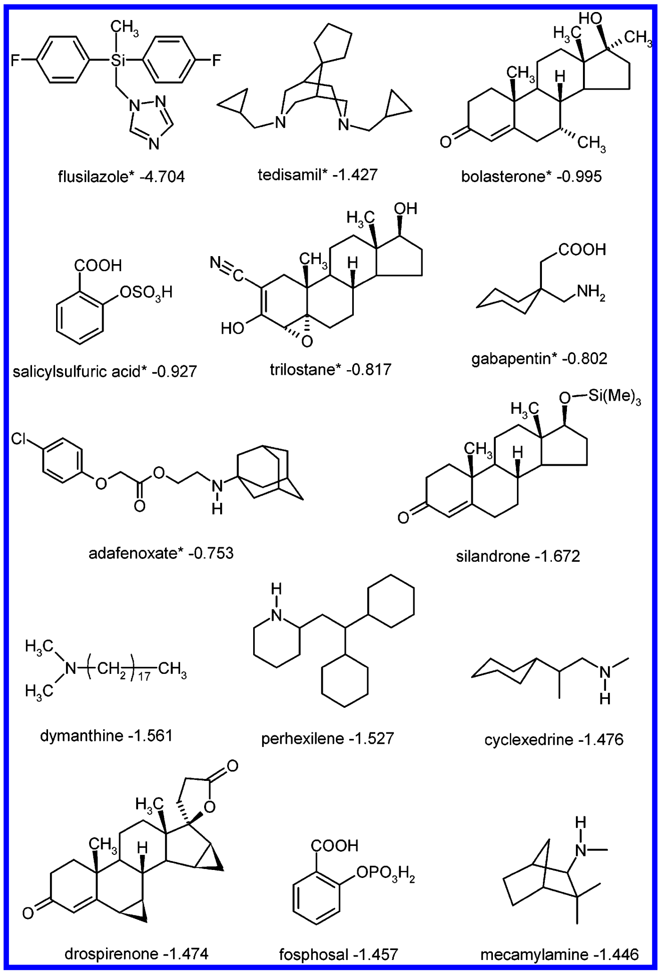

Figure 5. Examples of falsely classified drugs with particular low drug-likeliness scores. Large negative scores indicate an increasing probability of that compound being nondruglike. An asterisk denotes compounds that were part of the test set.

Table 3. Calculated Drug-Likeliness Score for a Number of Top-Selling Drugs[a]

names	core	name	score
ranitidine	1.383	lovastatin	−0.195
enalapril	0.635	cimetidine	0.654
fluoxetine	0.315	omeprazole	1.762
acyclovir	0.700	cefaclor	1.087
simvastatin	−0.209	ceftriaxone	1.429
amoxicillin	1.071	estrone	−0.201
clavulanic acid	1.109	equilin	0.155
diclofenac	0.342	cyclosporin	1.252
ciprofloxacin	2.501	beclometasone	−0.652
nifedipine	0.627	famotidine	0.980
captopril	1.226	salbutamol	0.014
diltiazem	0.904	sertaline	0.607
ezetimibe	0.429	imatinib	−0.190
celecoxib	1.199	losartan	0.478
sildenafil	0.692	loratadine	1.049
orlistat	−0.182	sibutramine	4.680

[a] List extended from ref 9.

that predict the druglikeness of compounds is somewhat difficult as the applied data sets and range of descriptors strongly vary. In many cases, descriptors such as logP are taken into account that contain physicochemical information about the molecules as a total entity. Here, solely atom types were used.

The best results in separating drugs from nondrugs are usually achieved by applying machine learning algorithms. The various neural network approaches of Sadowski and Kubinyi,[9] Ajay et al.,[10] Frimurer et al.,[11] Schneider et al.,[13,17] and Takaoka et al.[18] obtained accuracies between 77 and 90% for comprehensive data collections such as the World Drug Index. Similar or better accuracies were achieved with support vector machines and decision trees.[16,18,19] Murcia-Soler et al. and Gálvez et al. applied considerably smaller data sets and obtained similar results.[14,15] Frequently, it is observed that the removal of reactive substances and known toxic compounds, as well as such molecules with either very low or very high molecular weight, improves the prediction accuracy for drugs. After the removal of according agents, the prediction accuracy was found to increase to 46.4% for the nondrug test set. Doing so, however, implies chemical and medicinal knowledge as additional selection criteria on top of the mathematical fitting or training procedure that is the substantial part of a regression, neural networks, support vector machines, and decision trees. In contrast to the mentioned studies, the method applied in this work does not contain any corresponding fitting scheme, as it is solely based on statistical means. It is thus one of the key intentions to show alternative partitioning/scoring schemes that are independent from conventional training or fitting procedures.

4. CONCLUSION

It was found that atom types and their pair-wise combinations are distributed statistically differently between drugs and nondrugs. On the basis of these differences, a fast method was derived that is able to predict drugs with a higher accuracy than nondrugs. As similar trends were also found by using neural networks or decision trees, one may therefore speculate that there is apparently a significant amount of nondrugs that might possess druglike properties such as oral bioavailability that either have not been recognized or have not been screened yet. The resulting drug-likeliness score uses a separating margin of 0.0 that evolves from the underlying statistical approach. “Typical” drugs exhibit scores greater than 0.3. To cover about 84% of all drugs, however, molecules with negative scores as low as −0.2 should also be considered. Conversely, nondrugs with scores above 0.0 often contain rare or unwanted substitutions that can easily be ruled out by methods that incorporate medicinal knowledge. Because such filters are usually computationally more demanding, it is attractive to use the presented drug-likeliness score in the first place for the screening of extensive substance compilations.

Please note, a detailed list of the computed likeliness scores for the individual interactions for the compounds of the test sets, as well as the full list of compounds in the training sets, in addition to the PERL programs to generate the required matrices on the basis of a set of compounds using the atom types described in the method section and to calculate the corresponding likeliness score are available from the author upon request.

REFERENCES AND NOTES

(1) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

(2) 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.

(3) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(4) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095−5099.

(5) Wang, J.; Ramnarayan, K. Towards Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524−533.

(6) 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.

(7) 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.

(8) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.

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

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

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Gálvez, J.; de Julián-Ortiz, J. V.; García-Domenech, R. General Topological Patterns of Known Drugs. *J. Mol. Graphics Modell.* **2001**, *20*, 84−94.

(16) 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.

(17) 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.

(18) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoskikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemist’s Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1269−1275.

(19) 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.

(20) Daylight Chemical Information Systems Inc., Suite 550, Aliso Viejo, CA 92656. See http://www.daylight.com for full details of SMILES and SMARTS.

(21) Flower, D. R. DISSIM: A Program for the Analysis of Chemical Diversity. *J. Mol. Graph. Model.* **1998**, *16*, 239−253.

(22) 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.

(23) Muegge, I.; Heald, S. L.; Brittelli, D. Simple Selection Criteria for Drug-Like Chemical Matter. *J. Med. Chem.* **2001**, *44*, 1841−1846.

(24) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(25) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(26) 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.

(27) 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 Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.

(28) *The Merck Index*, 13th ed.; Merck & Co., Inc.: Whitehouse Station, NJ, 2001.

(29) Milne, G. W. A. *Drugs: Synonyms and Properties*, 2nd ed.; Asgate: Aldershot, Hampshire, U. K., 2000.

(30) *Janssen Chimica 88−90*; Janssen Pharmaceutica: Titusville, NJ, 1988.

(31) *HYPERCHEM*, version 6.02; Hypercube Inc.: Gainsville, FL, 1999.

(32) Karlin, S.; Altschul, S. F. Methods for Assessing the Statistical Significance of Molecular Sequence Features by Using General Scoring Schemes. *Proc. Natl. Acad. Sci. U. S. A.* **1990**, *87*, 2264−2268.

CI600329U