

A Large Descriptor Set and a Probabilistic Kernel-Based Classifier Significantly Improve Druglikeness Classification

Qingliang Li,^{†,‡} Andreas Bender,[§] Jianfeng Pei,[‡] and Luhua Lai^{*,†,‡}

Beijing National Laboratory for Molecular Sciences, State Key Laboratory of Structural Chemistry for Stable and Unstable Species, College of Chemistry and Molecular Engineering, and Center for Theoretical Biology, Peking University, 100871 Beijing, China, and Lead Discovery Center, Novartis Institutes for BioMedical Research Inc., 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received March 26, 2007

Probabilistic support vector machine (SVM) in combination with ECFP_4 (Extended Connectivity Fingerprints) were applied to establish a druglikeness filter for molecules. Here, the World Drug Index (WDI) and the Available Chemical Directory (ACD) were used as surrogates for druglike and nondruglike molecules, respectively. Compared with published methods using the same data sets, the classifier significantly improved the prediction accuracy, especially when using a larger data set of 341 601 compounds, which further pushed the correct classification rates up to 92.73%. On the other hand, most characteristic features for drugs and nondrugs found by the current method were visualized, which might be useful as guiding fragments for de novo drug design and fragment based drug design.

1. INTRODUCTION

High-throughput screening (HTS) and combinatorial chemistry have made significant impacts on the process of modern drug discovery.¹ As a large number of chemical compounds have been synthesized and screened in parallel, there is an increasing need for rapid and efficient models to profile druglike properties.² Although there are many reasons why potential drug candidates failed to reach the market, including toxicity, lack of efficacy, and market reasons, about 40% of the development compounds failed to reach the market due to poor pharmacokinetics, such as poor solubility, permeability, metabolic, and elimination profile.³ More recently, the contribution of off-target effects to adverse reactions⁴ and the influence of promiscuity⁵ have been examined more closely with respect to their contributions to cause failures in clinical development. In the prospective sense, to gauge whether a chemical compound is rather “druglike” or not as early as possible in the drug discovery pipeline will be extremely valuable.

While the concept of “druglikeness” may not be easy to define without a context such as the target, target class, and/or chemotype under consideration, many efforts have been made in this field in the past decade, and the term “druglike” has been used extensively. While different authors used this term slightly differently,^{6–9} Walter and co-workers⁶ summarized that “druglikeness” means that molecules contain functional groups and/or have physical properties consistent with the majority of known drugs. Furthermore, they

described the methods to recognize druglikeness including simple counting methods, functional group filters, chemistry space evaluation, and neural networks in the review. The other specific properties implicit in the general term of “druglikeness”, such as oral absorption, aqueous solubility, permeability, blood-brain barrier (BBB) penetration, and stability, were reviewed by Clark et al.,⁸ Blake et al.,⁹ and Di et al.¹⁰ Generally speaking, the former methods are usually applied to eliminate potentially unwanted molecules from compound libraries, while the latter ones that focus on the specific properties are normally used in lead optimization process.

Among various druglike predicting methods, probably the best-known one is the “rule of 5” developed by Lipinski¹¹ and co-workers at Pfizer, which was derived from an analysis of 2245 drugs from the World Drug Index (WDI).¹² It was found that poor absorption or permeation was more likely to happen when there were more than 5 H-bond donors, 10 H-bond acceptors, the molecular weight was larger than 500, and the calculated Log P (CLogP) was larger than 5. Unfortunately, Frimurer et al.¹³ found that the “rule of five” cannot discriminate compounds in the MACCS-II Drug Data Report (MDDR)¹⁴ from those in the Available Chemicals Directory (ACD),¹⁵ which are also typically considered as surrogates of druglike molecules and nondruglike molecules, respectively. Later, Muegge et al.⁷ employed a filter of pharmacophore points to discriminate druglike molecules from the Comprehensive Medicinal Chemistry (CMC) and nondruglike molecules from the ACD. In addition, Zheng et al. developed a chemical space filter using the molecular saturation related descriptor and the proportion of heteroatoms in a molecule to analyze the compounds from MDDR, CMC, ACD, and their own Chinese Natural Product Database (CNPD).¹⁶

* Corresponding author phone: (86)-010-62757486; fax: (86)-010-62751725; e-mail: lhlai@pku.edu.cn. Corresponding author address: College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China.

[†] College of Chemistry and Molecular Engineering, Peking University.

[‡] Center for Theoretical Biology, Peking University.

[§] Novartis Institutes for BioMedical Research Inc.

Currently, machine learning methods are also used in these applications. Using the WDI, CMC, and MDDR as surrogates for druglike molecules and the ACD as nondruglike molecules, Frimurer et al.,¹³ Sadowski and Kubinyi,¹⁷ and Ajay et al.¹⁸ constructed neural network models together with different molecular descriptors to classify druglike and nondruglike molecules. The various descriptors included one-dimensional parameters such as molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds, aromatics density, etc. and the Ghose-Crippen (GC) descriptors,¹⁹ two-dimensional parameters ISIS keys,²⁰ and numerically encode of 2D atom types. Byvatov and co-workers²¹ applied both support vector machine (SVM) and artificial neural network (ANN) methods using three distinct molecular descriptors including GC descriptors, descriptors provide by MOE software package (Molecular Operating Environment, Chemical Computing Group Inc., Montreal, Canada), and CATS topological pharmacophore descriptors.²² They concluded that using the complete set of all three descriptors, both SVM and ANN could get the best results, and SVM outperformed the ANN classifier with regard to overall prediction accuracy. Furthermore, Muller et al.²³ improved the prediction using the same data set with a careful model selection using support vector machines. Especially Wagener et al.²⁴ combined Ghose-Crippen descriptors with decision trees for the prediction of druglikeness, resulting in surprisingly simple rules. By just testing the presence of hydroxyl, secondary or tertiary amino, carboxyl, phenol, or enol groups, 75% of the drug data set was correctly predicted. Their initial decision tree even used the simple presence of hydroxyl groups as a predictor for membership of the group of drugs. Nondrugs, on the other hand, were found to be characterized by their aromatic nature, along with a low content of functional groups besides halogens.

Several reviews of the druglikeness classifications have been published recently which we would like to refer to for a more comprehensive coverage of literature on the topic.^{8,25,26} It could generally be observed that in many of the first studies such as the original work by Sadowski and Kubinyi¹⁷ as well as more recent studies such as the one by Muller et al.²³ descriptors were used which were not optimal in the sense that they were not able to generate different feature sets for compounds in the drug data set, as opposed to those in the nondrug data set. The often-employed Ghose-Crippen descriptors¹⁹ comprise 110 different atom-centered fragments, and they were originally used to model octanol–water partition coefficients, which represents a real-valued output variable. In order to avoid problems such as underdetermined equation systems, a rather smaller number of fragments are beneficial since the size of data sets is usually limited to at most a few thousand molecules. But as discussed previously,²³ this small number of features is in other cases not able to generate different feature vectors for compounds from the drug data set on the one side and compounds from the nondrug data set on the other side. More recently, the topological CATS descriptor was thus applied, able to discriminate between drugs and nondrugs a bit better than previous studies.²¹ In the current study, we employ a circular fingerprint descriptor which has previously shown to be a very well performing descriptor in virtual screening

studies,^{27–29} and we show in the following that this is transferable to drug/nondrug classification.

In this study we use the ECFP_4 (Extended Connectivity Fingerprints) as implemented in PipelinePilot 5.1 (SciTegic, Inc.) to characterize the molecules and employ a probability SVM model to classify druglike and nondruglike molecules. The results of discriminating molecules from the WDI and ACD as well as the comparisons with previously employed methods are summarized in the Results section, along with a visualization and discussion of the fragments most characteristic and least characteristic for compounds from both data sets.

2. DATA AND METHODS

2.1. Preparation of Data. In this study, the WDI and ACD (version 2002.2) were taken as surrogates for “druglike” and “nondruglike”, respectively. We preprocessed the two databases in the following way (in agreement with the work by Sadowski and Kubinyi¹⁷): (i) the records should contain valid connective table fields and no obvious errors; (ii) duplicates were removed in each individual database; and (iii) identical compounds found in both databases were removed from the ACD. For this purpose, the “Find Novel Molecules” function of PipelinePilot 5.1 was employed, which filters molecules based on identical canonical SMILES. As a result, we obtained 43 185 compounds from the WDI and 307 624 compounds from the ACD that could be well processed by our descriptor generation program.

Although ACD has been used widely in previous druglikeness studies,^{17,21,23} it may have a different physicochemical profile compared to the WDI set. In order to verify this, we have analyzed the physicochemical properties of both ACD and WDI (shown in Figure 1). All four properties were similar, among which the molecular weight distribution was very similar. Thus, we also used WDI and ACD for the current study.

Data set 1 and data set 2 have been constructed based on these two databases.

2.1.1. Data Set 1. Data set 1 was constructed in the same way as that in the previous work to allow comparison between the methods.^{21,23} We randomly extracted a subset of 9208 compounds from both the ACD and the WDI data sets as described above. With an 80 + 20 random split, the data set was divided into two parts, 80% of the samples for training and 20% for testing. This splitting step was repeated five times, corresponding to a 5-fold cross-validation. The above procedures were repeated ten times with random scrambling of molecules into a training and a test set, and an average measure of the performance for the classifier was obtained.

2.1.2. Data Set 2. As there were plenty of samples of “drugs” and “nondrugs” data in both databases, we also planned to train the classifier with a bigger data set, which was constructed in the following way. The whole data set from both the WDI and ACD excluding data set 1 was considered as a training set which contained 341 601 molecules. Then 80% of data set 1 was used to tune the parameters of the model, and the other 20% was for blind testing. For this training data set, the prior ratio (WDI/ACD = 1:8) of druglike and nondruglike molecules from the WDI and ACD was kept.

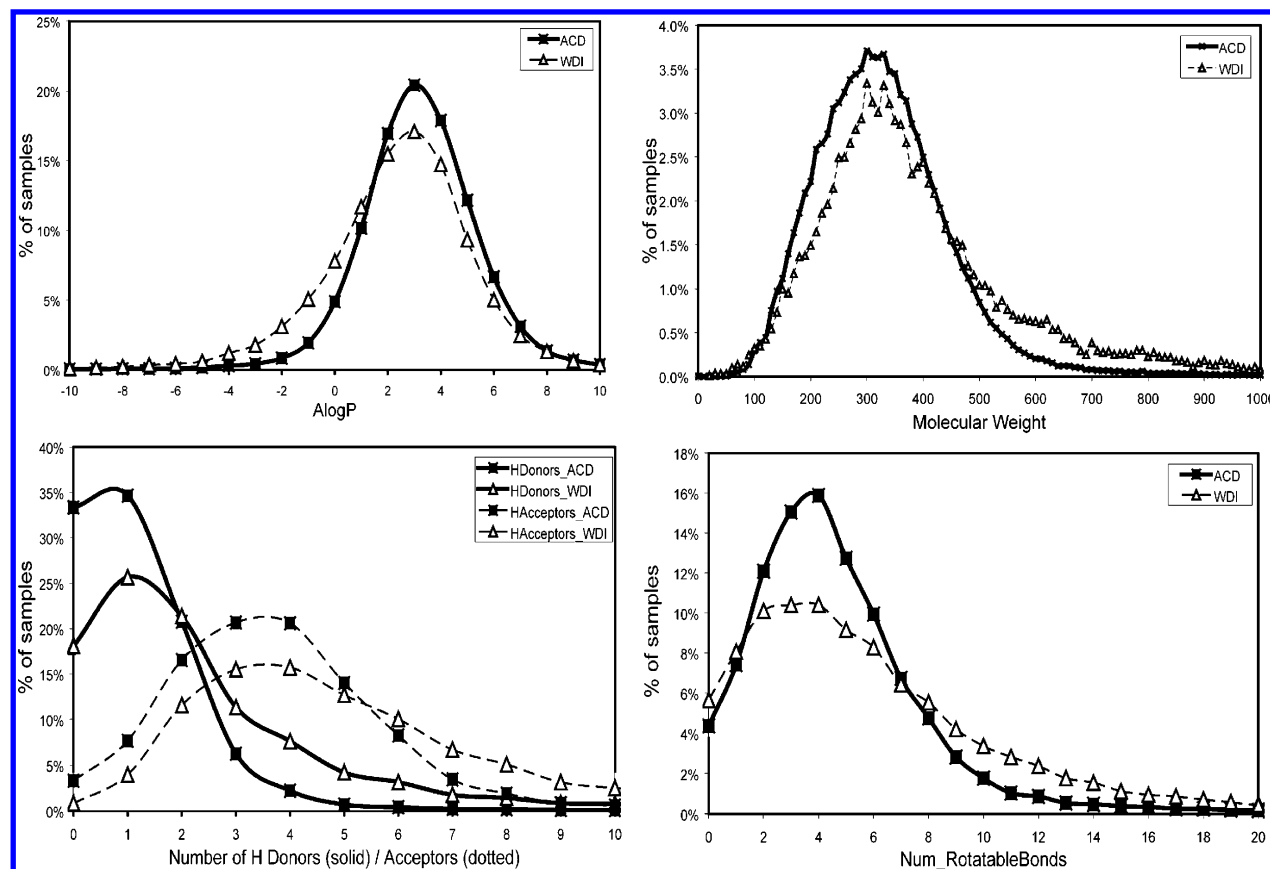


Figure 1. Property distributions of ACD and WDI data sets for AlogP, molecular weight, and the number of hydrogen bond donors and hydrogen bond acceptors as well as the number of rotatable bonds.

2.2. Molecular Descriptor. In this study we used the ECFP₄ (Extended Connectivity Fingerprints) as implemented in PipelinePilot 5.1 to characterize the molecules. It represents each fragment consisting of a heavy atom up to a diameter of 4 bonds in an integer format, and they have been shown to be of very good performance in virtual screening studies.²⁷ Atom types according to the number of bonds, the element type, the charge, and the mass are assigned. We also compared this descriptor to a different circular fingerprint descriptor, MOLPRINT 2D,^{30,31} which used different atom types (namely, SYBYL atom types³²), but it was found to give an inferior performance (see the Results section). For further information on ECFP₄³³ and MOLPRINT 2D circular fingerprints the reader is referred to the literature. The general construction of circular fingerprints is shown in Figure 2. For each heavy atom in turn, the atom types of the central atom and the atom types of its neighbors (and its neighbors of neighbors, etc.) are collected and put into a vector format. As mentioned above, the precise atom typing differs in each fingerprint definition.

2.3. Support Vector Machine. The support vector machine (SVM), a promising machine learning method first introduced by Vapnick and colleagues,³⁴ is based on the structural risk minimization principle from statistical learning theory. It is often considered superior to other supervised learning methods to some extent and has been employed in a wide range of fields.³⁵

In this work, we used the Libsvm³⁶ software package developed by Lin's group at the National Taiwan University. Although it contains several kernel functions including the

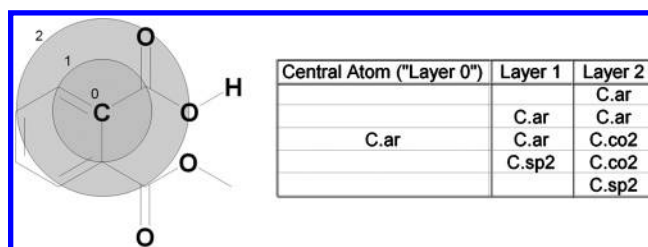


Figure 2. Illustration of the concept of circular fingerprints. For each heavy atom in turn, the atom types of the central atom and the atom types of its neighbors (and its neighbors of its neighbors...) are collected and put into a vector format. In the case illustrated here, mol2 force field atom types are employed. In case of ECFP₄ fingerprints the number of bonds, the element type, the charge, and the mass are assigned to each atom.

polynomial kernel, the sigmoid kernel, and the radial basis function (RBF), here we only employed the linear kernel and one nonlinear kernel function (RBF), as it provides good performances in most cases.^{23,37,38} On the other hand, instead of giving the prediction result of 'yes' or 'no', we applied a probability model of the SVM,³⁹ which can give a probability estimation of the testing molecules. In this way, we can pick out the druglike molecules according to a probability estimate. Moreover, the most characteristic features for druglike and nondruglike are calculated by linear support vector machines.

2.4. Model Evaluation. The performance quality of the model was examined with 5-fold cross-validation analysis. In the cross-validation test, the whole data set was shuffled, and then it was split 5-fold. Each fold was singled out in

Table 1. Statistical Data of the Model Evaluation for Linear and RBF Kernel^a

descriptor	method	data set size	accuracy (%)	sensitivity (%)	specificity (%)	Matthews cc
MOLPRINT 2D	SVM-Linear	9208	82.21 ± 0.77	84.81 ± 1.34	79.12 ± 0.86	0.641 ± 0.016
MOLPRINT 2D	SVM-RBF	9208	83.22 ± 0.82	86.78 ± 1.17	78.99 ± 0.81	0.662 ± 0.017
ECFP_4	SVM-Linear	9208	85.01 ± 0.24	86.07 ± 0.25	83.77 ± 0.59	0.697 ± 0.006
ECFP_4	SVM-RBF	9208	87.71 ± 0.26	88.63 ± 0.81	86.62 ± 0.42	0.753 ± 0.006
ECFP_4	SVM-Linear	341 601	89.74	89.90	89.55	0.79
ECFP_4	SVM-RBF	341 601	92.73	93.10	92.73	0.85

^a Each calculation was repeated three times independently.

turn for testing, and the remaining part ($n-1$ folds) of the data set was used for the training model. We used sensitivity (Q_p), specificity (Q_n), and overall accuracy (Q_a) to measure the accuracy of positive prediction (compounds belong to druglike data set), negative prediction (compounds belong to nondruglike data set), and the overall accuracy of the model, respectively, like previous studies.^{40,41}

$$Q_p = TP/(TP + FN)$$

$$Q_n = TN/(TN + FP)$$

$$Q_a = (TP + TN)/(TP + TN + FP + FN)$$

Here, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. In general, the overall accuracy of Q_a is always used to measure the prediction power of a model. For comparison we also calculated the Matthews correlation coefficient (cc),^{40,41} which ranges from -1 to 1 and where a perfect prediction gives a correlation coefficient value of 1 .

$$cc = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}}$$

This correlation coefficient (cc) is suitable for measuring the prediction power for an unbalanced data set (where the numbers of positive and negative samples are not equal).

3. RESULTS AND DISCUSSION

3.1. Results of Data Set 1. Data set 1 consists of the same size of druglike and nondruglike molecules and an identical ratio of training and testing sets as those of the reported methods. The results are shown in Table 1, and the distribution of the predicted scores is plotted in Figure 3A/B.

Preliminary studies comparing different descriptor types were undertaken to establish descriptor performance. Namely, MOLPRINT 2D circular fingerprints and ECFP_4 fingerprints were compared, which differ with respect to the atom types used as well as the fact that MOLPRINT 2D used *only* fingerprints with a given radius, while ECFP_4 fingerprints also include all fragments of smaller diameter which can be constructed from a given molecule (results are given in Table 1). It can be seen that ECFP_4 fingerprints gave a consistently superior performance compared to MOLPRINT 2D on all four performance measures (accuracy, sensitivity, specificity, and Matthews correlation coefficient). Therefore, only ECFP_4 fingerprints were employed in further studies.

3.1.1. Linear-SVM Model. First, we trained an SVM employing the linear kernel function (linear-SVM) on data set 1. The only parameter, C , of the linear SVM was tuned

to optimize the model, and 5-fold cross-validation was used to evaluate the performance. On average, the linear SVM model yields an accuracy of 85.01% with the correlated coefficient of 0.70. The distribution of the predicted scores is shown in Figure 3A. White bars and black bars represent the predicted results of the ACD and WDI test data sets, respectively. We can see that most of the predicted scores of the nondruglike compounds from the ACD are lower than 0.2, while the majority of the predicted scores of druglike molecules from the WDI are higher than 0.8. Fewer predicted scores of the molecules from both ACD and WDI distributed in the region between 0.4 and 0.6. Therefore, a borderline according to the scoring value can be easily drawn to separate druglike and nondruglike compounds.

3.1.2. RBF-SVM Model. Furthermore, we applied a second SVM with a RBF kernel to data set 1. Two parameters of RBF-SVM, the kernel parameter γ and the regulatory constant of C , were regulated during the training process. Here we used the grid.py tool in the Libsvm package to systematically search the parameter space in order to get the best model. The 5-fold cross-validation results are shown in Table 1, and the distribution of the predicted scores is shown in Figure 3B.

A slight improvement in the overall accuracy and the correlation coefficient was compared to the counterpart of results by the linear model (Table 1). An accuracy of 87.71% and a correlation coefficient of 0.75 are obtained when employing an RBF kernel, compared to 85.01% and a correlation coefficient of 0.70 for the linear kernel. More than half of the druglike compounds scored higher than 0.9, most of the nondruglike compounds scored less than 0.2, and fewer compounds were distributed in the region of 0.3–0.7 in the histogram, compared to the results shown in Figure 3A.

The results of two models above indicate support vector machine can give a good performance in druglike classification, and the discriminative power of the nonlinear model (RBF-SVM) is superior to linear model (linear-SVM) in this study, which is also confirmed by other work.^{18,28,29} Furthermore, it is noticeable that we employed a probabilistic of an SVM model,³⁹ which gives a probability estimate of class memberships. According to different probability scores, one can effectively pick out a series of compounds in a certain confidence from a chemical library.

3.2. Results of Data Set 2. Given the large size of both ACD and WDI, two additional SVM models with linear and RBF kernels were trained on data set 2. The results are also shown in Table 1, and the distributions of the probability scores are plotted in Figure 3C/D.

3.2.1. Linear SVM Model. We followed a similar training strategy to train data set 2 as that for data set 1. Since the

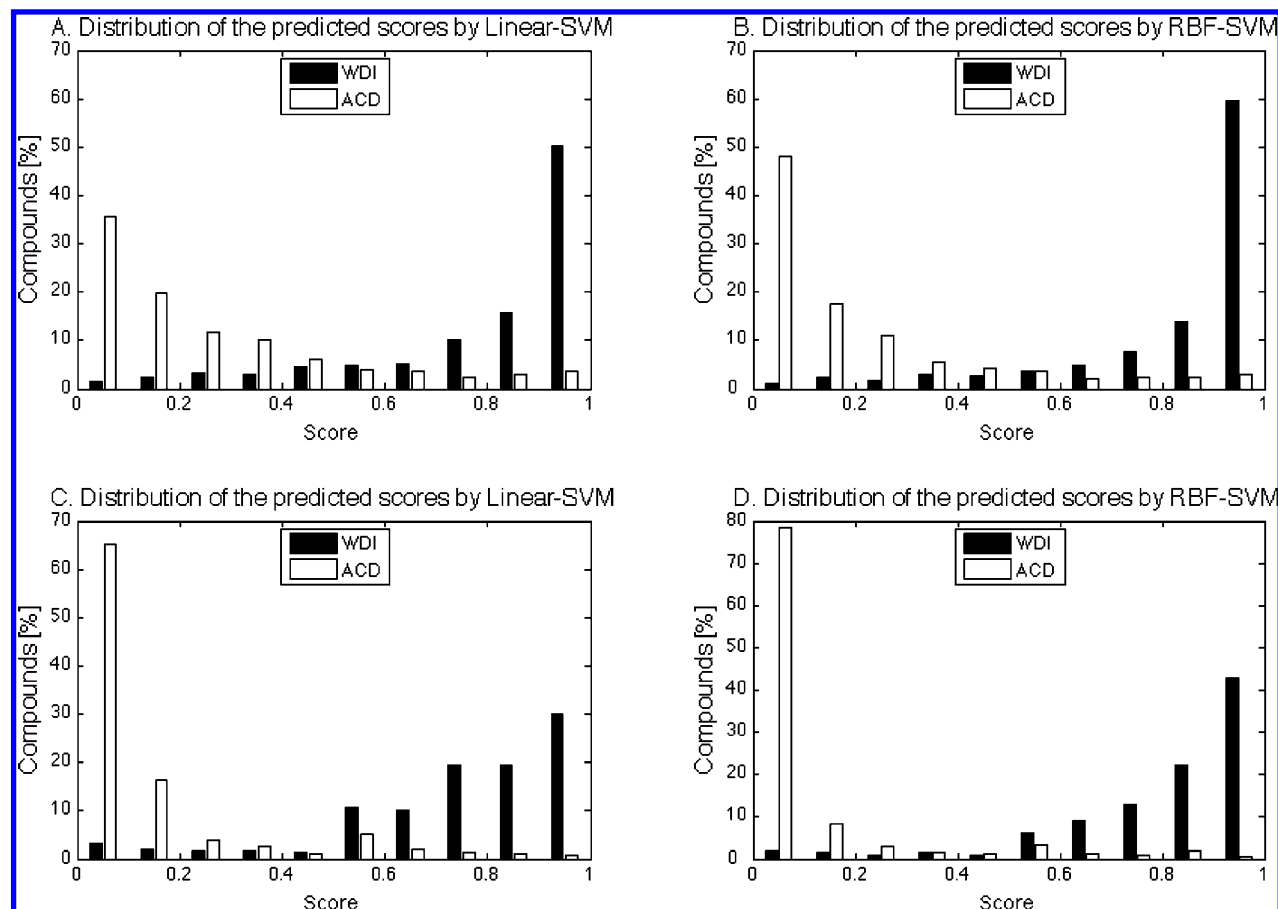


Figure 3. Histograms to illustrate the distribution of predicted probability scores: **A.** distribution of scores predicted by Linear-SVM trained by data set 1; **B.** distribution of scores predicted by RBF-SVM trained by data set 1; **C.** distribution of scores predicted by Linear-SVM trained by data set 2; and **D.** distribution of scores predicted by RBF-SVM trained by data set 2.

size of data set 2 is much larger than that of data set 1, to train a SVM with such a huge data set is an extremely time-consuming process. Instead of using cross-validation, we used 80% of the left molecules for training and 20% for testing. From Table 1, we can observe about 4% improvement in the overall accuracy (to 89.74%) compared to the results of the linear SVM trained on the smaller data set 1, and 1% improvement over the RBF SVM trained on the smaller data set. The Matthews correlation coefficient is 0.79. The distribution of the predicted scores is shown in Figure 3C, where a clear separation of druglike compounds from nondruglike compounds can be observed.

3.2.2. RBF SVM Model. Furthermore, the RBF-SVM was also applied to the larger data set 2. The results of this model are better than those of all the models above. The best results (accuracy of 92.73%, $cc = 0.85$) were achieved with a sensitivity of 93.10% and the specificity of 92.73%, respectively. These numbers are about 3% higher in overall accuracy than that of the linear-SVM with the larger data set 2, and 5% better than that of the RBF-SVM with the smaller data set.

Generally (as can be seen in Table 1), we can observe a clear improvement in performance with the increasing data set size, independent of the kernel used (linear or nonlinear kernel). This might be due to either better sampling of the “drug” and the “nondrug” space or, assuming that the global druglikeness model consists of a multitude of individual local models, due to better sampling of those “drug islands”. On

Table 2. Comparison of the Results Reported Here with Previously Published Methods

no.	descriptors	methods	data set size	% correct	Matthews cc
1 ^a	GC	SVM	9208	80.01 ± 0.087 (81.9 ± 0.4^b)	0.592 ± 0.002
2 ^a	MOE	SVM	9208	80.19 ± 0.74	0.593 ± 0.016
3 ^a	CATS_225	SVM	9208	73.90 ± 0.51	0.485 ± 0.011
4	ECFP_4	SVM	9208	87.71 ± 0.26	0.753 ± 0.006
5	ECFP_4	SVM	341 601	92.73	0.85

^a Results of nos. 1–3 come from ref 21. ^b This result comes from ref 23, while the Matthews cc is not available in the reference.

the other hand, as the number of compounds in the ACD is much larger than the ones in WDI, the original compound ratio of ACD and WDI is also kept in this study.²³

3.4. Comparison to Other Methods. The comparison to other methods is summarized in Table 2. In order to make results comparable, we constructed training data set 1 in the same way as it was done in previous work.^{21,23}

As reported before,²¹ for Ghose-Crippen and MOE descriptors, between 80.01% and 81.9% correct classification rates were obtained, while the CATS model yielded slightly worse results (73.9%). Our study using ECFP_4 descriptors and RBF-SVM classifiers gave a correct classification rate of 87.71% which achieves about 6–7% improvement. When training with the larger data set of 341 601 compounds, classification accuracy as high as 89.74% and 92.73% was obtained using a linear and an RBF kernel, respectively.

Muller et al.²³ used a data set of around 186 000 compounds and achieved a classification accuracy of 89.2%; by using a large data set of 341 601 compounds, we improved the classification accuracy to more than 92% (though not strictly comparable due to the different data set sizes). This implies that both appropriate molecular descriptors and a large enough data set are crucial for the correctness of classification.

Overall, we are able to increase classification performance for each of the data sets and kernels studied, which thus corroborates the suitability of the ECFP_4 descriptor to reflect the differences of compounds from the ACD and WDI databases; a concept we here also entitle “druglikeness”. (This might in turn either be a global model or a set of multiple, local druglikeness models for different targets, target classes, or chemotypes, as discussed in section 3.8.) As far as descriptor choice is concerned, we again find that a suitable description of the problem under consideration is crucial for the performance that can be expected from the model. The fundamental problem of the machine learning method is how to characterize samples with precise and informative features. The ECFP_4 fingerprint descriptor was employed to characterize molecules in a druglikeness study due to its well representation of the structure information. For the ECFP_4 fingerprint, each heavy atom in a molecule was assigned to a different atom type according to its physicochemical property. And the atom type and the type of its neighbors (and its neighbors of neighbors—up to a diameter of 4 chemical bonds) were extracted to form a fragment in which the local information of the different part of the molecule was stored. As a molecule it consisted of many kinds of individual fragments, and in this way most of the characteristics of the whole molecule were gathered. The efficiency of the ECFP_4 fingerprint in druglikeness classification of this study was also in agreement with similar findings in the context of compound classification⁴² and in the realm of target prediction for small molecules.^{43,44} Another circular molecular descriptor, the MOLPRINT 2D fingerprint,^{30,31} outperformed previously reported descriptors, which demonstrates that such kind of descriptors (circular descriptor) can effectively represent the characteristics of molecules.

Although the fingerprints of MOLPRINT 2D and ECFP_4 are all circular descriptors, they gave different performances in druglikeness classification. There might be two reasons why the MOLPRINT 2D fingerprint did not perform as well as the ECFP_4 fingerprint did—the MOLPRINT 2D fingerprint in combination with RBF SVM achieved around 80% classification accuracy. One is that instead of using the SYBYL atom type,³² the ECFP_4 fingerprint assigned each atom type according to its physicochemical properties, and the other is that ECFP_4 used a set of circular fingerprints up to a given size, while MOLPRINT 2D took only the ones with a fixed diameter into account.

3.6. Misclassified Compounds. The misclassified compounds can be distinguished into false negative compounds (drugs which are falsely predicted to be nondrugs) as well as false positive compounds (nondrugs which are falsely predicted to be drugs). The 10 structures from each set assigned to the wrong category with the highest probability are shown in Table 3 (false negatives) as well as in Table 4 (false positives).

Table 3. Structures from the World Drug Index with Low Druglikeness Scores

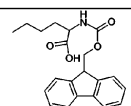
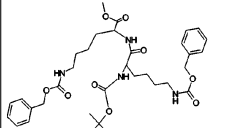
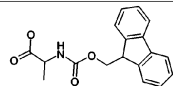
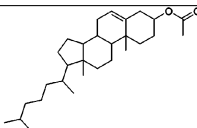

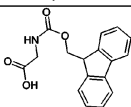
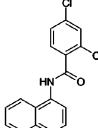
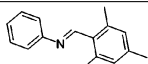
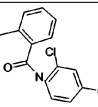
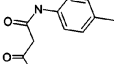
No.	Probability Score	WDI False-Negative Structures
1	0.00006	
2	0.00008	
3	0.012	
4	0.012	
5	0.013	
6	0.019	
7	0.020	
8	0.026	
9	0.026	
10	0.027	

Table 3 shows the set of false negative compounds, listing structures taken from the World Drug Index which are assigned to the lowest druglikeness scores. Overall it can be seen that the structures are, on average, rather lipophilic and possess only a small number of typical druglike functional groups (such as alcohols or amines). This is in agreement with earlier work, which even led to a druglikeness filter simply based on the number of pharmacophoric points per molecule, which performed surprisingly well, given its simplicity (classifying about two-thirds of both drugs and nondrugs correctly).²⁵ On the other hand, some of them actually contain very druglike scaffolds, such as structure 4 which possesses the steroid core. Actually there are number of compounds in ACD containing steroid scaffold indeed, which might be one reason. Another reason is probably due to their small number of hetero atoms, relative to the large carbon steroid core structure. Especially as has previously been found, steroids are likely to be misclassified by other methods as well.⁴⁵

Table 4 shows the set of false positive compounds, listing structures from the Available Chemicals Directory, which are assigned the highest druglikeness scores. It can be seen

Table 4. Structures from the Available Chemicals Directory with the Highest Druglikeness Scores

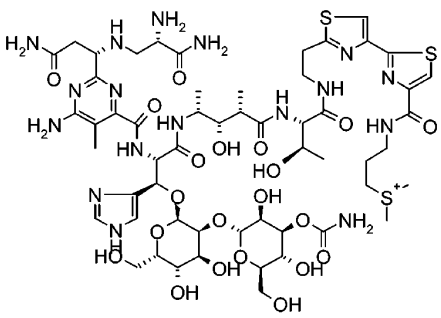
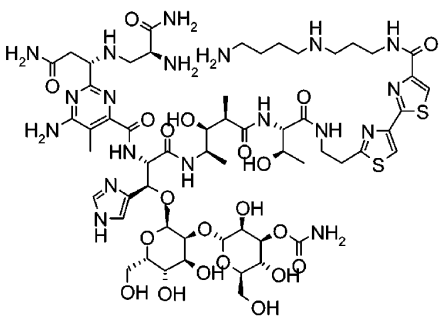
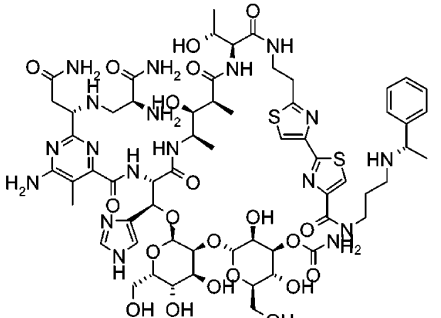
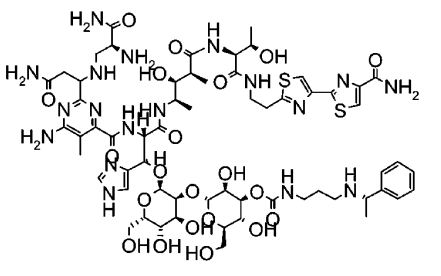
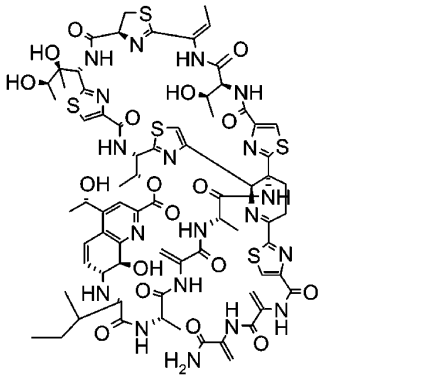
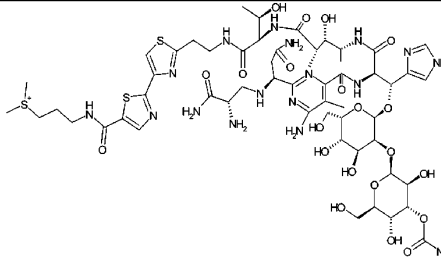
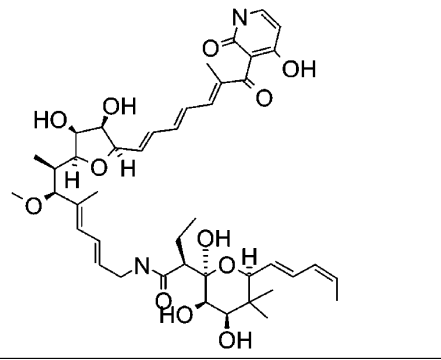
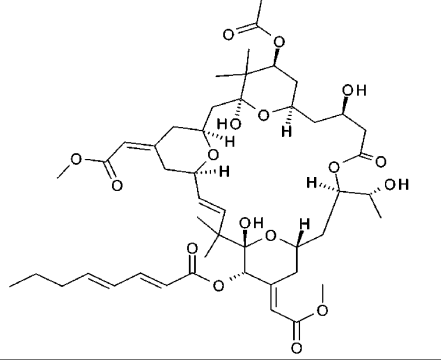
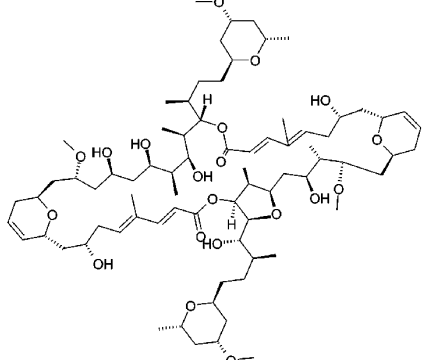
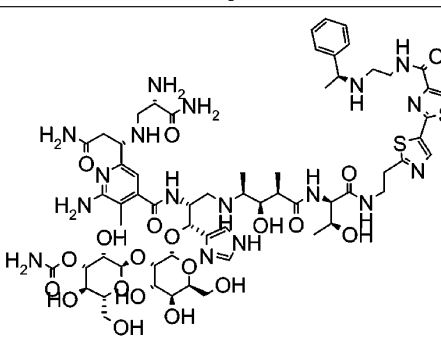
No.	ACD ID	Probability Score	ACD False-Positive Molecules
1	MFCD01862603	0.999	
2	MFCD01740724	0.999	
3	MFCD01708900	0.999	
4	MFCD01076537	0.999	
5	MFCD00135828	0.999	

Table 4 (Continued)

No.	ACD ID	Probability Score	ACD False-Positive Molecules
6	MFCD00070310	0.999	
7	MFCD00467139	0.999	
8	MFCD00893832	0.999	
9	MFCD03095588	0.999	
10	MFCD00468087	0.999	





















Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability
	1.155	-0.7476		0.7433	1.316		0.6887	0.4740		0.6405	1.640
	0.9080	0.7759		0.7381	0.6300		0.6628	0.6720		0.6141	1.231
	0.8211	0.8484		0.7055	1.679		0.6607	1.048		0.6141	1.315
	0.7674	1.000		0.7030	1.496		0.6541	0.6714		0.6009	1.137
	0.7666	1.290		0.6947	0.9563		0.6524	0.2336		0.5940	1.441

Figure 4. The 20 structural features statistically associated with the drug data set, according to the weight in the linear model. Shown are linear SVM weights along with normalized probabilities from a Bayes model. Features range from 0 to 2 bonds in diameter, and fragments are protonated according to physiological pH.

that the structures are, on average, rather large and possess a large number of typical druglike functional groups (sugar moieties, amides, alcohols) which makes them resemble the “typical” natural products. After further analysis it was found that those top 10 “false-positive” compounds are indeed natural products used as antibiotics—1, 2, and 6 are bleomycin derivatives, structures 3 and 4 are peplomycins, structure 5 is thioestrepton, structure 7 is “antibiotic MYC-8003”, structure 8 is bryostatin 1, and compound 9 is isoswinholide A. Thus, we are tempted to conclude that the exclusion of drugs from the ACD “nondrug” data set was not perfect, due to either slight variations of the structures present or the algorithm used to eliminate duplicate structures via canonical SMILES. While unable to make statements about previous work, depending on the methods used it might have been possible that a similar situation was encountered by other groups.

3.7. Druglike Features. The 20 features statistically most associated (largest weights) with the drug data set as well as those features least associated with the drug data set are shown in Figure 4 (drug features) and Figure 5 (nondrug features).

In addition to the weight in the linear SVM model, Bayes normalized probabilities are listed. To be clearly understood, we would like to simply explain the definition of the normalized probability. The naïve Bayes is a statistical classification method based on the Bayes rule of conditional probability that given A compounds are drugs in total T compounds (drugs + nondrugs), feature F_i is contained in T_{F_i} compounds and A_{F_i} compounds containing feature F_i are drugs. Then the druglikeness estimate of a molecule containing feature F_i should be $P(\text{druglike}|F_i) = A_{F_i}/T_{F_i}$. To avoid the overconfident estimate when a feature F_i is small, we employed a Laplacian corrected estimator^{46,47} which samples the feature T/A additional times. Therefore the final probability is $P_{\text{final}}(\text{druglike}|F_i) = (A_{F_i} + 1)/(T_{F_i} (A/T) +$

1]. For features more common in drugs $\log P_{\text{final}} > 0$, while for features less common in drugs $\log P_{\text{final}} < 0$. But in some cases, the normalized probabilities differ considerably from the weights assigned by the SVM model. Features range from 0 to 2 bonds in diameter, and fragments are protonated according to physiological pH. It should be noted that ECFP_4 fragments include all fragments of diameter of four and below, thus smaller fragments are also present in the model.

Wagener et al.²⁴ employed a classification tree to simply test the presence of hydroxyl, tertiary or secondary amino, carboxyl, phenol, or enol groups and can distinguish drugs and nondrugs at a success rate of about 75%. That hydroxyl groups are very characteristic for drugs is in agreement with earlier work, which is shown here both in the top feature in Figure 4 as well as the feature in position 2 and in two instances showing phenol rings. The second most characteristic feature was trialkyl substituted nitrogens, here featured at the top of the right-hand column in Figure 4. The third most characteristic feature is Al_2NH moieties, here again featured in two instances, one in the diethyl amine version (right-hand column) as well as the dimethyl amine moiety (third column of Figure 4). Thus, good agreement with earlier features can be stated, while at the same time it is obvious that in the current work larger fragments are employed for classification.

Compared to Hutter,⁴⁵ we also observed that aromatic rings are more characteristic of nondrugs than drugs (10 instances in Figure 5 vs 6 instances in Figure 4) and that sp^3 and sp^2 hybridized nitrogens are more common among drugs. Interestingly, Hutter observes that alcoholic, ether, and ester oxygens prevail in nondrugs, which is the opposite of result of the work discussed above and also of our results.

It should be noted that the features shown here contribute *in their entirety* to the druglikeness classification of a compound, so each feature only gives a small contribution

Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability	Molecule	Score	NormalizedProbability
	-0.1922	-1.313		-0.1925	-3.707		-0.1931	-0.6472		-0.1937	1.103
	-0.1922	-1.849		-0.1926	-2.014		-0.1932	-2.146		-0.1937	-1.655
	-0.1924	-1.662		-0.1928	0.5319		-0.1932	0.4430		-0.1938	-1.495
	-0.1925	0.7161		-0.1929	-3.146		-0.1933	-1.118		-0.1938	0.3748
	-0.1925	-2.430		-0.1929	1.083		-0.1936	-0.6739		-0.1938	-0.2832

Figure 5. The 20 structural features statistically associated with the nondrug data set, according to the weight in the linear model. Shown are linear SVM weights along with normalized probabilities from a Bayes model. Features range from 0 to 2 bonds in diameter, and fragments are protonated according to physiological pH.

(depending on the fragment weight) to the overall score, and no single feature is sufficient for classification. A recent example demonstrating this characteristic of cumulating scores of individual features has been published for HIV protease inhibitors in combination with the Bayes classifier.⁴⁸ Additionally, these meaningful features might be useful as guiding fragments in fragment based drug design or de novo drug design to improve the druglike property.

3.8. Discussion of the Concept of “Druglikeness”. In the present study, we were able to improve the current benchmark of druglikeness prediction by employing a more fine-grained descriptor set. For applications of this method in drug discovery, we should nonetheless be aware of its conceptual limitations.

One of the limitations of this model concerns its structural applicability domain. Some compound classes such as antibiotics or vitamins are generally not within the scope of these models. Strict application to filter out nondruglike compounds may thus narrow the scope for future discoveries. While taking the druglikeness prediction as one of many indicators, it would be hazardous to eliminate compounds purely based on this consideration.

Databases add another layer of uncertainty. While we know for sure that some compounds possess druglike properties (those which made it at least to the clinical stages), we hardly ever know about the “proven nondruglikeness” of a compound. Even if a compound was found to be ineffective in one disease, e.g., due to poor solubility, there might be another receptor against which it is much more active, and where thus solubility problems are not an issue anymore.

In addition, as discussed in a very recent publication on the topic,⁴⁹ large numbers of analogue compounds and questions of generalizability from one activity class to another should be kept in mind. And if there are two databases for which we would like to assign class member-

ship are given—what does our model predict? Is it the underlying property we are attempting to model, here druglikeness? Or maybe it is only that particular subset of drugs our databases contain—so it is rather “World-Drug-Indexlikeness” or “current-known-druglikeness” our model detects? Finally, the chemistry of the training sets is naturally limited to the chemistry presently known (and publicly available), which is a common limitation of models in general and restricts its predictivity with respect to novel chemotypes of “chemistry” in general.

Still, also given the limitations discussed above we were able to show that we can improve upon current classification schemes by different descriptor selection, of descriptors with sufficient information capacity for the problem under investigation. In any case, we should not expect perfect classification, due to the reasons mentioned in this section.

4. CONCLUSION

In this study we have developed a druglikeness classifier using an extended connectivity fingerprints (ECFP_4) descriptor with probability support vector machines. Employing previously used data sets, the performance of the classifier could be improved from 80.19% (MOE descriptors with SVM) or 81.9% (Ghose-Crippen descriptors with SVM) to 87.71% for the combination presented here. SVM with RBF kernels perform better than with a linear kernel, yielding accuracies of 85.01% and 87.71%, respectively. When using a much larger data set containing 341 601 compounds, performance of the SVM model was significantly increased to over 90% (92.73% with an RBF kernel and 89.74% with a linear kernel). Here we conclude that a suitable description of the molecules and the larger data set are crucial for achieving good performance in druglikeness classification in this study. On the other hand, most characteristic features for drugs and nondrugs were found by the current method,

which might be useful as guiding fragments for fragment based drug design and de novo drug design.

ACKNOWLEDGMENT

This work was supported by the State Key Program of Basic Research of China (2003CB715900) and the High-Tech Program of China. A.B. thanks the Education Office of Novartis for a Postdoctoral Fellowship.

REFERENCES AND NOTES

- Entzeroth, M. Emerging trends in high-throughput screening. *Curr. Opin. Pharmacol.* **2003**, *3*, 522–9.
- Egan, W. J.; Walters, W. P.; Murcko, M. A. Guiding molecules towards drug-likeness. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 540–549.
- Venkatesh, S.; Lipper, R. A. Role of the development scientist in compound lead selection and optimization. *J. Pharm. Sci.* **2000**, *89*, 145–54.
- Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007**, *2*, 861–873.
- Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling Promiscuity Based on in vitro Safety Pharmacology Profiling Data. *ChemMedChem* **2007**, *2*, 874–880.
- Walters, W. P.; Murcko, M. A. Prediction of 'drug-likeness'. *Adv. Drug. Delivery. Rev.* **2002**, *54*, 255–271.
- Muegge, I. Selection criteria for drug-like compounds. *Med. Res. Rev.* **2003**, *23*, 302–21.
- Clark, Pickett, Computational methods for the prediction of 'drug-likeness'. *Drug Discovery Today* **2000**, *5*, 49–58.
- Blake, J. F. Chemoinformatics - predicting the physicochemical properties of 'drug-like' molecules. *Curr. Opin. Biotechnol.* **2000**, *11*, 104–7.
- Di, L.; Kerns, E. H. Profiling drug-like properties in discovery research. *Curr. Opin. Chem. Biol.* **2003**, *7*, 402–8.
- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.
- World Drug Index (WDI). Derwent Information: London, U.K.
- Frimurer, T. M.; Bywater, R.; Naerum, 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.
- MACCS-II Drug Data Report (MDDR). Molecular Design Limited: San Leandro, CA.
- Available Chemicals Directory (ACD). Molecular Design Limited: San Leandro, CA.
- 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*, 856–62.
- Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325–3329.
- Ajay, A.; Walters, W. P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314–24.
- Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for 3-Dimensional Structure-Directed Quantitative Structure-Activity-Relationships. 1. Partition-Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, *7*, 565–577.
- SSKEYS; MDL Information Systems Inc.: San Leandro, CA.
- 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.
- Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894–2896.
- Muller, K. R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'drug-likeness' with Kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249–53.
- Wagener, M.; van Geerestein, V. J. Potential drugs and nondrugs: prediction and identification of important structural features. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 280–92.
- Muegge, I.; Heald, S. L.; Brittelli, D. Simple selection criteria for drug-like chemical matter. *J. Med. Chem.* **2001**, *44*, 1841–1846.
- Walters, W. P.; Ajay, Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3*, 384–7.
- Glen, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. *IDrugs* **2006**, *9*, 199–204.
- Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOL-PRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
- Fichert, T.; Yazdani, M.; Proudfoot, J. R. A structure-permeability study of small drug-like molecules. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 719–22.
- Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
- Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOL-PRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–18.
- SYBYL Atom Types; Tripos Inc.: 1699 South Hanley Road, St. Louis, MO 63144, U.S.A.
- Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.
- Cortes, C.; Vapnik, V. Support-vector networks. *Machine Learning* **1995**, *20*, 273–297.
- Xu, Y.; Zomer, S.; Brereton, R. G. Support Vector Machines: A recent method for classification in chemometrics. *Crit. Rev. Anal. Chem.* **2006**, *36*, 177–188.
- Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines; 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Ding, C. H.; Dubchak, I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* **2001**, *17*, 349–58.
- Dobson, P. D.; Doig, A. J. Distinguishing enzyme structures from non-enzymes without alignments. *J. Mol. Biol.* **2003**, *330*, 771–83.
- Ting-Fan, W.; Chih-Jen, L.; Ruby, C. W. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. Mach. Learn. Res.* **2004**, *5*, 975–1005.
- Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16*, 412–424.
- Lapinsch, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. *Biochim. Biophys. Acta* **2001**, *1525*, 180–190.
- Cannon, E. O.; Bender, A.; Palmer, D. S.; Mitchell, J. B. Chemoinformatics-based classification of prohibited substances employed for doping in sport. *J. Chem. Inf. Model.* **2006**, *46*, 2369–80.
- Jenkins, J. L.; Bender, A.; Davies, J. W. In silico target fishing: Predicting biological targets from chemical structure. *Drug Discovery Today: Technologies* **2006**, *3*, 413–421.
- Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging chemical and biological space: "target fishing" using 2D and 3D molecular descriptors. *J. Med. Chem.* **2006**, *49*, 6802–10.
- Hutter, M. C. Separating drugs from nondrugs: a statistical approach using atom pair distributions. *J. Chem. Inf. Model.* **2007**, *47*, 186–94.
- Nidhi; Glick, M.; Davies, J.; Jenkins, J. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–33.
- Xia, X.; Maliski, E.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463–70.
- Klon, A. E.; Glick, M.; Davies, J. W. Application of machine learning to improve the results of high-throughput docking against the HIV-1 protease. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2216–2224.
- Llinas, A.; Burley, J. C.; Box, K. J.; Glen, R. C.; Goodman, J. M. Diclofenac solubility: independent determination of the intrinsic solubility of three crystal forms. *J. Med. Chem.* **2007**, *50*, 979–83.