

Can We Learn To Distinguish between “Drug-like” and “Nondrug-like” Molecules?

Ajay,* W. Patrick Walters, and Mark A. Murcko

Vertex Pharmaceuticals Inc., 130 Waverly Street, Cambridge, Massachusetts 02139

Received October 3, 1997

We have used a Bayesian neural network to distinguish between drugs and nondrugs. For this purpose, the CMC acts as a surrogate for drug-like molecules while the ACD is a surrogate for nondrug-like molecules. This task is performed by using two different set of 1D and 2D parameters. The 1D parameters contain information about the entire molecule like the molecular weight and the the 2D parameters contain information about specific functional groups within the molecule. Our best results predict correctly on over 90% of the compounds in the CMC while classifying about 10% of the molecules in the ACD as drug-like. Excellent generalization ability is shown by the models in that roughly 80% of the molecules in the MDDR are classified as drug-like. We propose to use the models to design combinatorial libraries. In a computer experiment on generating a drug-like library of size 100 from a set of 10 000 molecules we obtain *at least* a 3 or 4 order of magnitude improvement over random methods. The neighborhoods defined by our models are not similar to the ones generated by standard Tanimoto similarity calculations. Therefore, new and different information is being generated by our models, and so it can supplement standard diversity approaches to library design.

1. Introduction

Recent developments in combinatorial organic synthesis and high throughput screening methods have enormously increased the possibility of finding novel lead compounds.^{1–3} The rate of discovery of novel leads depends, in part, on the diversity of the assayed compounds. To this end there have been many attempts to define and construct diverse sets of compounds so that screening can be made more effective than random search.^{4–7} Many reasonable arguments have been made to show that a designed library is preferable to a random one. However, this is still an act of faith. The primary reason is that it is not obvious that diversity is independent of the system being assayed, that is, a diverse database for one biological target may not be equally diverse for another. In addition, when measuring diversity it is not obvious how to weight the descriptors that describe a molecule. The major hurdle currently faced by researchers is the lack of sufficient experimental results to evaluate different questions on diversity.

Combinatorial libraries can be truly enormous. Even virtual libraries created on a computer have to be truncated in order to evaluate their characteristics. A simple and useful example is to limit the analysis to synthetically accessible molecules. We propose the notion of “drug-likeness” as another useful way to select molecules for screening. We contend that there is value in designing a library (either diverse or focused) which contains a set of “drug-like” compounds. The ability to distinguish drug-like molecules will also be useful for other computational chemistry endeavors like *de novo* design.⁸ Therefore, it is important to understand and evaluate the concept of drug-likeness of a molecule. In this paper we describe a first attempt at the task of

predicting drug-likeness. We perform this task by using a set of 2D/1D descriptors to train a learning system.

What makes a molecule a drug? The answer depends on a complicated interrelationship between toxicity, synthetic accessibility, chemical and metabolic stability, cost, marketability, and so on. There have been previous attempts at predicting toxicity,^{9,10} synthetic accessibility, and metabolic stability. These ventures have seen limited success. The prediction of what constitutes a drug should be proportionately more difficult. What provides us hope is the availability of large databases of drug or drug-like molecules, e.g., CMC (Comprehensive Medicinal Chemistry),¹¹ and MDDR (MACCS-II Drug Data Report).¹² Large datasets are not readily accessible in other, more specialized, areas like toxicity or metabolism. In addition we also have much larger databases of compounds represented by the ACD (Available Chemicals Directory).¹³ We view the molecules in the ACD as falling into two categories. They are either (1) “close-to-drugs” or (2) “far-from-drugs”. So if we would like to build a small library from the ACD we would be well-advised to pick up a set of diverse compounds that are close-to-drugs.

There are over 80 000 compounds in the CMC and MDDR databases together, with over 5000 compounds in the CMC after we have eliminated compounds like spermicides, aerosol propellants, etc.¹⁴ We need to develop methods that will extract useful information from such large databases. These methods for *data mining* will either enumerate patterns from or fit models to data. There are many statistical problems with blind data mining that we need to be careful about. For example, large databases can be a blessing against overfitting. On the other hand, we have to guard against chance fits especially in systems that search through large model spaces (large number of descrip-

* Author to whom correspondence should be addressed. Tel: 617 577 6000. Fax: 617 577 6400. E-mail: ajay@vpharm.com.

tors). High dimensionality of the model space (i.e., many descriptors and functional forms) also causes problems as our understanding of estimation in high dimensions is fairly primitive. Another major problem with large databases is that they grow over time and not always as if sampled from a static probability distribution, i.e., the probability distribution can alter drastically. For example, Lipinski¹⁵ has shown that there has been a shift to higher molecular weights for the compounds in clinical trials over the past few years. That is, the characteristics of drug molecules today may change in the future. The simplest solution to this problem is retraining as new data arrive.

2. Methods

In this section we provide an overview of the learning systems used to analyze the CMC and ACD databases. Additional details can be found in the Supporting Information.

2.1. The Learning Systems. As mentioned in the Introduction, data mining can be fraught with chance fits. To guard against this, we have used two completely different procedures to build the models, one based on neural networks¹⁶ and the second on a machine learning algorithm (c4.5)¹⁷ which is a decision tree learning system. A neural network stores its information in a distributed fashion and small changes in parameters and input values does not affect its performance. On the other hand, c4.5 is rather more sensitive to the magnitude of input values. Therefore, the learning characteristics of these algorithms are entirely different, providing a buffer against chance fits. Both the methods require some training data from which learning takes place. As the size of the ACD and CMC are large, we have chosen to work with smaller subsets (this is computationally required for both the neural networks and part of the c4.5 suite of programs). The training set is formed by a random partition of about 3500 compounds each from the CMC and the ACD (this leaves about 2000 compounds for testing from the CMC, a sufficiently large number to gauge predictive performance.). Ten different training/test set realizations were constructed and studied to minimize the possibility of chance fits. In addition we have conducted experiments with random data to test the efficiency and usefulness of the models constructed.

We note an important distinction between the standard way learning systems are used and their use here. The learning system is usually asked to pick a model based on the descriptors that best distinguish between two (in our case) classes of compounds, drugs and nondrugs. The model generated is then used for gaining insight or for prediction. During training and testing the classes assigned to the compounds are always correct—there is no ambiguity in this assignment. In our case, however, the learning system will be trained based on the assumption that the vast majority of compounds in the ACD are nondrugs. After training, we hope to then evaluate compounds in the ACD as drug-like or nondrug-like based on the prediction by the model and its level of certainty about the prediction. That is, we are interested in teasing out drug-like compounds from the set that were initially classified as nondrugs. We will therefore be relying solely on the system's assessment of its correctness for predictive purposes.

The basic intuition behind this paper is described in Figure 1. It shows the distribution of objects belonging to two classes (circles and squares) in a two-dimensional space. It also shows two decision boundaries (curved lines) generated by two hypothetical learning algorithms. A good learning algorithm will choose to generate a decision boundary that separates the two classes as much as possible. Note that the squares found in regions of space predominantly occupied by circles are more "circle-like" than other squares. Squares near the boundaries separating the classes are intermediate in nature. If we assume that the circles represent compounds from the CMC and the squares compounds from the ACD, we would like a set of descriptors (and appropriate learning methods) that

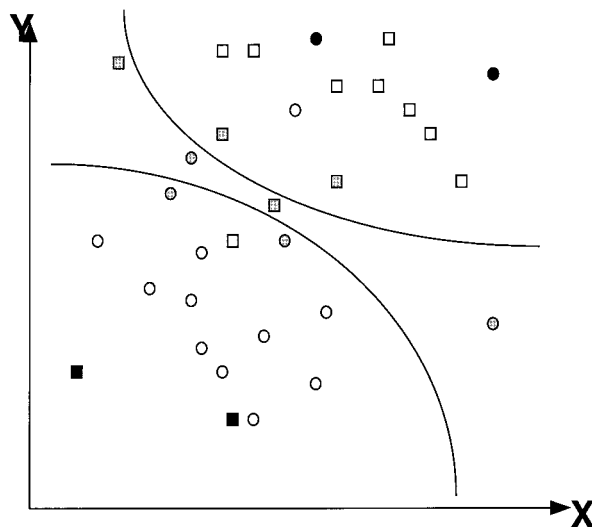


Figure 1. Basic intuition. The curved lines shows two possible decision boundaries. The filled squares or circles are misclassified with low probability while the shaded ones are misclassified with high probability. See text for interpretation.

would put most of the circles on one side of the decision boundary and as many squares as possible on the other side. Once such a decision boundary is generated we can pick the squares which were misclassified with low probability (i.e., squares found in regions of descriptor space dominated by circles) and call them more drug-like than others. In general, the distribution of the points in space will depend on the descriptors used and so will our ability to draw decision boundaries. Also, the filled circles and squares represent objects that are misclassified with low probability (low confidence) and the shaded ones are misclassified with high probability (high confidence).

In the rest of the section we will describe our choice of descriptors for the molecules and provide a rough outline of Bayesian neural networks (BNN). Details regarding the both c4.5 and BNN algorithms can be found in the Supporting Information.

2.1.1. Descriptors. The choice of descriptors is difficult a priori. We cannot hope to use thousands of descriptors in a learning system. We also want to limit the computational resources used in building descriptors. These issues have led us to the choice of 2D and 1D descriptors. The 1D ("one-dimensional") descriptors contain information about the entire molecule. As a first test, we started with a set of seven descriptors,¹⁸ namely $\log P$, molecular weight (MW), number of hydrogen bond donors (ND), number of hydrogen bond acceptors (NA), number of rotatable bonds (NR), aromatic density (AR), and the kappa index (${}^2\kappa_a$, specifying the degree of branching of the bonding pattern;¹⁹ in general this will be highly correlated with NR). Intuitively, these seem a reasonable first choice. The 2D ("two-dimensional") parameters contain information about the presence or absence of specific functional groups within a molecule. Our choice for this set is based on the ISIS fingerprint²⁰ for each compound. This is a bit string (a string of 0's and 1's) of length 166 with a 1/0 indicating the presence/absence of some moiety or "key". There are 166 such keys for each compound (Table 1 lists some of these). This choice of descriptors is reasonably common in the diversity literature.²¹ In fact, Brown and Martin²¹ have recently shown that the 166 ISIS keys perform remarkably well for clustering and diversity analysis.

Building reliable models with the ISIS fingerprints is much harder than with the seven descriptors because of the high dimensionality. Prediction accuracy, however, should improve with the ISIS fingerprints as it is capable of representing molecules in greater detail.

It is possible to graphically explore the distribution of the seven descriptor set. This would allow us to look for outliers

Table 1. Description of Each of the ISIS Keys^a

1*,†,#	isotope	86*,†,#	CH2QCH2
2*,†,#	103 < at.no < 256	88*,#	S
3*,†,#	group IVA,VA,VIA etc.	89*,†,#	OAAAO
8*,†,#	QAAA@1	93†	QCH3
9*,†,#	group VIII; metals	96*,#	five-membered ring
12*	group IB, and IIB element	99*,†	C=C
13*,†,#	ON(C)C	100*,#	ACH2N
15*	OC(O)O	101†	eight-membered or larger ring
17*,†	C#C	103#	Cl
18#	group IIIA element	105*,†,#	ASA(SA)\$ A
19*,†,#	seven-membered ring	106*,#	QA(Q)Q
20*,#	Si	112*,#	AA(A)(A)A
22*,†	three-membered ring	116*	CH3AACH2A
23#	NC(O)O	121*,†	heterocycle with N
25†,#	N-O	125*,#	more than one aromatic ring
26*,#	CS=C(SA)SA (all bonds ring bonds)	129*,†	ACH2AACH2A
30*,#	fragment CQ(C)(C)A	131†,#	more than one QH group
31†,#	QX	135*,†	Nnot%A%A
35*,#	group IA element	137†,#	heterocycle
36#	sulfur-containing heterocycle	138*,†,#	OH
38*,#	NC(C)N	143*	ASA!O
39†	OS(O)N	144†	Anot%A%Anot%A
42†,#	F	145*,#	more than one six-membered ring
44*,†,#	other (?)	147*,†	ACH2CH2A
47*,†,#	SAN	148*,#	AQ(A)A
52*,†,#	NN	151†	NH
54*	QHAAQH	152*	OC(C)C
55*,†	OSO	158*,†,#	C-N
56*,†,#	ON(O)C	162#	aromatic
60*,†	S=O	163*,#	six-membered ring
62*,†,#	ASA!ASA	166*,#	more than one fragment; structure cannot be drawn in one connected set
64*,#	AS-!AS		
65*,†,#	C%N; aromatic bond	167*,†,#	MW
66*	CC(C)(C)A	168*,†,#	donor
68*,†,#	QHQH	169*,#	acceptors
72*,†	OAAO	170*,#	rotors
77*,#	NAN	171*,†,#	aromatic density
80*,†	NAAAN	172*,#	² K _α
82*	ACH2QH	173*,†,#	log P
84*,†,#	NH2		

^a “=” is a double bond, “#” is a triple bond, “Q” is any heteroatom, “A” is any heavy atom, “X” is a halogen, “%” is for ring bond, “not%” for not ring bond, “\$” is an aromatic bond, “not\$” is not an aromatic bond, “!” means the bond must be part of a chain. “Hn” means at least *n* hydrogens must be present. The “@*n*” indicates that atom number *n* is attached here. Parentheses indicate branching (connected to the left atom, but not the right). If the bond is omitted, then it means any type of bond. The “+ rare” represents the presence of some other rare features in the molecule. The “starred” keys are important for the linear models (BNN0). The keys with a “†” are important in providing nonlinear contributions in the networks with five hidden units (BNN5). Finally the ones with important linear contributions in BNN5 are designated with a pound (#) sign. There is a significant overlap between the BNN0 keys and the linear contributors from BNN5, though there are some differences.

and perhaps eliminate them from consideration. The reason for removing compounds with very large or very small values (outliers) is not because we deem them unimportant. Rather, outlying values may inhibit the model construction process. In addition we are interested in constructing general purpose databases, and we would like to eliminate compounds with peculiar properties. No such graphical analysis, however, is possible for the ISIS fingerprints.

With the ISIS fingerprint data it is possible to have a fair number of unimportant descriptors. This can also cause problems with any learning method. We have conducted limited investigations on retraining after removing the unimportant descriptors within the Bayesian neural network framework. The major advantage of such pruning in descriptor space accrues from the fact that a more compact decision space usually leads to more robust classifiers. Robust classifiers are in general more reliable for predictions.

2.1.2. Bayesian Neural Network. Neural networks can be viewed as a flexible regression (classification) technique. Because of the inherent nonlinearity of neural networks they are able to model regularities in data much more effectively than linear models. The almost unlimited “extra flexibility” provided by a neural network often causes modeling of not just the regularities but stray correlations in the data. There are many ad hoc procedures that attempt to prevent learning of these chance correlations. The most important of these is the use of weight decay and early stopping procedures based on

monitoring the performance on an independent validation set.¹⁶ Sometimes, cross-validation is also used to monitor and evaluate performance. However, cross-validation estimates can often be quite noisy leading to difficulties in parameter optimization. A Bayesian approach to neural network modeling^{22,23} allows for simultaneous and reliable optimization of a large number of control parameters. It allows us to sample the weight space more thoroughly than standard methods, thus avoiding local minima pitfalls.

In standard neural network training we are interested in obtaining a single set of weights that fit the training data, i.e., minimize an error criterion. In contrast within a Bayesian procedure we obtain a large number of weights and associated with each set is a probability weighting factor that is high if the resulting error is low and vice versa.

It is intuitively clear that most choices of weights will lead to low probability weightings. It is therefore imperative that a reliable and robust method be found for sampling in weight space. This is done in analogy with accepted procedures in molecular dynamics and Monte Carlo methods used in protein simulations. Technically these methods are called Markov Chain Monte Carlo (MCMC) methods.²⁴ In this paper we have used the hybrid MCMC procedure advocated by Neal.²² More details of the algorithm are described in the appendix.

Since this is a binary classification problem we use a network with 1 output unit. The standard Gaussian data model for linear regression would be inappropriate and so is

replaced by a logistic regression model.¹⁶ We have built networks with no hidden units (BNN0; this is analogous to linear discriminant analysis), a network with five hidden units (BNN5), and one with ten hidden units (BNN10).

2.1.3. Testing Protocols. It is nontrivial to construct fool-proof criteria to assess the success of the learning system on our problem. The reasons for this include the large dimensionality of the problem and relative size of the ACD database compared to the CMC.

We use five criteria to assess reliability: (1) the accuracy and confidence in prediction of compounds; (2) consistency or otherwise of predictions by the different methods; (3) an examination of the change in predictions when small changes are made to the compounds; (4) a visual examination of drug molecules that are incorrectly predicted; and (5) behavior of predictions upon randomization of both descriptors and classes.

We also report on the classification accuracy on two different databases: (1) MDDR, (2) a small set of 30 compounds built from scaffolds¹⁴ and side chains²⁸ generally found in known drugs from the CMC database. This explores predictive behavior on small changes to a molecule. These results should provide a good assessment of the practical usefulness of our procedure. Exploring the performance on the MDDR database provides an opportunity to assess extrapolations. One way to assess the usefulness of the method is by looking at the number of compounds in the MDDR that are classified as drug-like. One would expect, a priori, that this number should be large, since the MDDR contains molecules believed to be biologically active.

3. Results

We begin with a discussion of the differences between the ACD and the CMC based on the seven-descriptor set. This is followed by detailed results based on the Bayesian neural network (BNN) method. The c4.5 results are somewhat worse (in all cases) than the BNN ones. They are more prone to local minima problems. Further details regarding the performance of the c4.5 algorithm can be found in the Supporting Information.

3.1. Differences in the Distribution of the Seven-Descriptor Set. Histogram plots of the seven-descriptor dataset do not reveal any differences between compounds in the CMC and ACD. A similar analysis looking at quantiles and QQ plots²⁵ does not reveal any differences either. The QQ plots, however, show that there are some outliers in the databases. We have therefore elected to prune the CMC and ACD to get rid of compounds with molecular weights greater than 600, with $\log P$ values beyond 11.0 and -4.0 , and with ${}^2\kappa_\alpha$ values larger than 30.²⁶ This was done so as to keep the models generated from becoming too biased.

The correlation matrix and principal component analysis (PCA) for the CMC and ACD databases are shown in Tables 2 and 3, respectively. Notice that even though the pattern of the correlations are similar there are significant differences in values. For the CMC, MW is correlated with (NA, NR, ${}^2\kappa_\alpha$) (coefficient > 0.5) while for the ACD, MW is correlated with (NR, ${}^2\kappa_\alpha$, AR, $\log P$). A principal component analysis (on correlations, i.e., standardized data) shows that the first three components explain roughly 84% of the data in both CMC and ACD. A factor analysis²⁷ (with three components) using the varimax method shows some differences between the CMC and the ACD. For the ACD, the first factor has large loadings on NR and ${}^2\kappa_\alpha$ and intermediate loadings on MW and $\log P$, the second factor has large loadings on ND, NA, and $\log P$, the third factor has large loadings on AR, MW, and $\log P$. For the CMC, the first

Table 2. Basic Statistics on the Compounds from the Pruned CMC Database

	Correlation						
	MW	ND	NA	NR	${}^2\kappa_\alpha$	AR	$\log P$
MW	1.0	0.13	0.58	0.55	0.73	0.31	0.36
ND		1.00	0.38	0.16	0.13	-0.16	-0.49
NA			1.00	0.35	0.38	-0.05	-0.37
NR				1.00	0.89	0.12	0.23
${}^2\kappa_\alpha$					1.00	0.22	0.39
AR						1.00	0.48
$\log P$							1.00
	Principal Components						
Eigenvalue	2.97	2.01	0.81	0.63	0.39	0.10	0.06
Percent	42.51	28.78	11.68	9.08	5.57	1.44	0.90
CumPercent	42.51	71.29	82.98	92.07	97.64	99.09	100.00
	Rotated Factor Patterns Assuming Three Factors Are Important						
MW		0.73		-0.24			0.46
ND		0.08		-0.78			-0.11
NA		0.39		-0.78			0.13
NR		0.93		-0.05			-0.05
${}^2\kappa_\alpha$		0.96		-0.03			0.13
AR		0.07		0.15			0.92
$\log P$		0.38		0.73			0.43

Table 3. Basic Statistics on the Compounds from the Pruned ACD Database

	Correlation						
	MW	ND	NA	NR	${}^2\kappa_\alpha$	AR	$\log P$
MW	1.00	0.13	0.48	0.56	0.68	0.57	0.55
ND		1.00	0.37	0.23	0.23	-0.05	-0.40
NA			1.00	0.35	0.38	0.13	-0.22
NR				1.00	0.86	0.14	0.29
${}^2\kappa_\alpha$					1.00	0.27	0.43
AR						1.00	0.54
$\log P$							1.00
	Principal Components						
Eigenvalue	3.18	1.81	0.91	0.59	0.27	0.12	0.10
Percent	45.49	25.83	12.99	8.44	3.96	1.82	1.44
CumPercent	45.49	71.32	84.32	92.76	96.73	98.55	100.00
	Rotated Factor Patterns Assuming Three Factors Are Important						
MW		0.56		-0.20			0.71
ND		0.16		-0.79			-0.13
NA		0.23		-0.79			0.26
NR		0.94		-0.15			0.05
${}^2\kappa_\alpha$		0.92		-0.15			0.23
AR		0.03		0.02			0.93
$\log P$		0.42		0.60			0.60

factor has large loadings on NR, ${}^2\kappa_\alpha$, and MW, the second on ND, NA, and $\log P$, the third on AR with intermediate ones on MW and $\log P$. It appears, therefore, that there is some difference between the two datasets at least with the seven-descriptor set.

A similar analysis was performed in order to look for biases introduced when considering only about 2% of the data from the ACD for training. Table 4 shows the correlation matrix and principal component analysis for the ACD compounds in one training set. In this case the numerical values for the correlation matrix, PCA, and factor loadings are very close to the complete ACD, indicating that the diversity of the ACD is captured to a significant degree by a dataset of roughly 4000 compounds. A similar statement can be made for all the 10 training sets. The same 10 train/test set pairs were used for all the results in this paper. We caution, however, that a reasonable coverage based on the seven-descriptor set does not necessarily imply a similar conclusion when using the ISIS fingerprints.

Table 4. Basic Statistics on the Compounds from the ACD in One of the Training Sets

	Correlation						
	MW	ND	NA	NR	$^2\kappa_\alpha$	AR	$\log P$
MW	1.00	0.14	0.47	0.55	0.68	0.57	0.55
ND		1.00	0.38	0.25	0.27	-0.05	-0.40
NA			1.00	0.36	0.39	0.13	-0.23
NR				1.00	0.87	0.13	0.28
$^2\kappa_\alpha$					1.00	0.26	0.39
AR						1.00	0.54
$\log P$							1.00
	Principal Components						
Eigenvalue	3.16	1.83	0.91	0.59	0.28	0.12	0.10
Percent	45.17	26.10	12.95	8.45	4.02	1.79	1.50
CumPercent	45.17	71.27	84.22	92.67	96.70	98.49	100.0
	Rotated Factor Patterns Assuming Three Factors Are Important						
MW		0.55		-0.19			0.72
ND		0.18		-0.78			-0.13
NA		0.23		-0.79			0.26
NR		0.94		-0.15			0.06
$^2\kappa_\alpha$		0.92		-0.17			0.24
AR		0.02		0.01			0.93
$\log P$		0.40		0.61			0.61

Table 5. Prediction Performance on 10 Independent Realizations of the Test Set Based on a Bayesian Neural Network with One Hidden Layer Containing Five Units (BNN5)^a

method	CMC error	ACD error	MDDR "drug-like"
BNN5; 7des	16–19	25–29	61–68
BNN5; ISIS	21–23	17–19	83–84
BNN5; ISIS+7des	9–11	11–12	77–79

^a The CMC database lists the percent misclassifications. The ACD also lists the percent misclassifications, in other words it gives the percent of drug-like compounds in the ACD. For the MDDR we list the percent of drug-like molecules. See the text for a discussion.

3.2. Prediction Performance. Table 5 shows the result when using a neural network with five hidden units trained using the Bayesian method. For the CMC and ACD databases the table reports the percent error made in classification; for the MDDR it reports the percentage of compounds that were classified as drug-like. The range of values shown in the table correspond to all the 10 networks.

The performance using just the seven-descriptor set is quite impressive. It predicts over 80% of the CMC compounds correctly. However, about 30% of the ACD is classified as drug-like, which appears to be high. A respectable number of compounds in the MDDR are classified as drug-like. This conforms with our expectation that any method that distinguished drug-like molecules from nondrug-like molecules should classify a majority of the MDDR compounds as drug-like.

Using 166 binary descriptors instead of the seven 1D descriptors results in a somewhat improved performance, overall. The error in the CMC test compounds is larger by about 5%, while number of compounds classified as drug-like in the ACD falls sharply by 10% and rises sharply by about 17% for the MDDR.

As expected, the performance of the networks that combine the ISIS keys and the seven-descriptor set is the best. It covers about 90% of the CMC space while only classifying about 10% of the ACD as drug-like and classifying around 80% of the compounds in MDDR as drug-like. The improvement in the predictive performance over the seven-descriptor set, especially the

Table 6. Prediction Errors on 10 Independent Realizations of the Test Set Using the 173 Descriptor Dataset for a Linear Discriminant Network (Using Bayesian Methods) and Network with 10 Hidden Units^a

method	CMC error	ACD error	MDDR "drug-like"
BNN0; ISIS+7des	14–17	16–17	76–77
BNN10; ISIS+7des	9–11	11–12	77–79

^a The results are shown in the same format as in Table 5.

Table 7. Performance Using the Reduced Set of 78 Descriptors on the CMC, ACD, and MDDR Databases^a

method	CMC error	ACD error	MDDR "drug-like"
BNN5; ISIS+7des (71 ISIS + 7des)	10–12	12–13	77–80

^a Notice that the performance falls negligibly from the networks using 173 descriptors. This shows the advantages of the Bayesian procedure.

reduction in the number of compounds in the ACD that are classified as drug-like, implies that we need to use the additional complication introduced by the ISIS keys.

Table 6 gives the predictive performance obtained by using a linear classifier (BNN0) and a network with 10 hidden units (BNN10). There are no differences in predictive performance between BNN5 and BNN10 to within the level of accuracy shown in the table. The performance of BNN0 is not as good as BNN5. The improvement in performance of BNN5 is significant enough to use BNN5 models for any future predictions.

The Bayesian learning procedure allows us to assign an "importance level" to each descriptor for generating a model. An analysis of the values of the hyperparameters associated with the input-hidden and input-output weights can tell us the importance of each descriptor to the model (see Supporting Information for details). We base the importance level of a descriptor on the median value of the hyperparameter associated with the input-output or input-hidden weights for the descriptor. The hyperparameter values are positive numbers, and in our case the maximum is usually below 20. All descriptors with median hyperparameter value less than or equal to 0.1 are deemed irrelevant. Next, all those descriptors that are considered to be important in at least 6 out of the 10 trained networks are reported in Table 1 (a total of 78 descriptors). The "starred" keys are important for the linear models (BNN0). Notice that all seven 1D descriptors are important. The keys with a "†" are important in providing nonlinear contributions in the networks with five hidden units (BNN5). Finally the ones with important linear contributions in BNN5 are designated with a pound (#) sign. There is a significant overlap between the BNN0 keys and the linear contributors from BNN5, though there are some differences. These differences arise due to the fact that BNN5 also has nonlinear contributions.

As a final check, Table 7 shows the results of training 10 networks (on the same training sets) using the reduced set of 78 descriptors. As expected, these results are close to the ones obtained using the original data. These results are a little worse than the complete set of 173 keys; this is expected as the hyperparameters removed from consideration were not completely irrelevant—their significance level was low compared to the other descriptors.

3.3. Randomization Tests. One of the final tests we performed was a data randomization test. The descriptors in the training and test sets were randomly permuted. Upon training, no learning took place—all the weight vectors were close to zero, implying a classification of “nondrug” for all compounds. This suggests that the learning methods are picking up some real information from the original unrandomized data.

Two additional, more stringent tests were performed in order to determine that there are real differences between drug-like and nondrug-like molecules and that some of this is captured by the CMC and ACD databases and the representation we have adopted. We constructed random training sets out of the ACD with half of the data labeled as “drugs” and other half as “acd”. The same experiment was conducted with the CMC database. We report on the results of training a Bayesian neural net on these two datasets. At this point it is important to recall that ISIS fingerprints are designed so that a very large percentage of compounds in the datasets (CMC, ACD, and others) are distinguishable. This implies that, in principle and unlike the descriptor randomization results, one would expect that a reasonable separation of any two random datasets is possible.

On the dataset that splits the ACD database into two sets, one called “drugs” and the other “acd”, we obtained the following results using BNN5.

1. The training set has ~35% error in the “acd” group and ~30% error in the “drug” group. This is very different from the ~15% and ~9% error rates in the real dataset.

2. CMC classification error is ~62%, versus 9% on the real data.

3. Approximately 80% of the molecules in the MDDR database were classified as nondrugs, versus 20% for the real data.

On the dataset that splits the CMC database into two sets, we obtained the following results using BNN5.

1. The training set has ~30% error in the “acd” group and ~28% error in the “drugs” group.

2. Approximately 50% of the compounds in the CMC test set were classified as “acd”, and ~56% of compounds in the ACD were classified as “drugs”.

3. Approximately 70% of compounds in the MDDR database were classified as non-drugs.

Both of these results indicate that our analysis has picked up some real differences between molecules in the CMC and the ones in the ACD.

3.4. Exploring Predictions on a Small Set of Compounds. In order to explore the consequences on predictions by making small changes in compounds, we include the results on a small list of 30 compounds given in Figure 2. This is a list generated by Guy Bemis from his work on elucidating the main features (frameworks and side chains) most often found in drugs.^{14,28} New molecules were generated by randomly combining a few frameworks with a few side chains. Figure 2 lists the classifications obtained from the complete list of 173 descriptors and the smaller set of 78 descriptors. The 173 descriptor set classifies 10 (out of 30) compounds as drug-like while the 78 descriptor set classifies 17 compounds as drug-like. In general from this small list, it appears that small changes to a compound does not generally change the classification. But, as the ex-

amples of benzoic acid to benzoic acid amide and 3,4-methylenedioxyphenol (sesamol) to 3,4-methylenedioxyaniline show, the classification is sensitive to small context-dependent changes. One curious result is the change in classification of aspirin from nondrug-like to drug-like in going from the 173 descriptor set to the 78 descriptor set.

Figure 2 also illustrates the difficulty encountered in trying to capture drug-like features in simple rules. For example, looking through the list of CMC compounds that the network misclassifies (compounds not shown), it is not easy to see any patterns. The same situation results when we explore the nondrug-like compounds in the MDDR (compounds not shown). In a sense this is reassuring as we would not expect any “simple” rules that will cover a vast majority of compounds in the CMC and MDDR.

3.5. An Experiment in Designing a Combinatorial Library. Consider a situation where we have a list of 10 000 compounds that we could purchase from a vendor. We would like to buy and assay only 1% of the compounds. Which 100 should we buy? If we have no information about the classes of molecules that will produce a hit in our assay, then we could simply choose the top 100 drug-like compounds. To study this situation in detail we consider sets of 10 000 compounds randomly drawn from the ACD. New sets are formed by randomly replacing 1, 5, 10, 50, 100, and 500 compounds from the original set by compounds randomly chosen from the CMC. Using the trained networks predictions are made (using the 173 descriptor set) on all 10 000 compounds and the top scoring 100 drug-like molecules are chosen from the list. In order to account for sampling error each such experiment is run 20 times. The results are given in Table 8. Given a set of a drug-like molecules in a set of size n , the average number of drug-like molecules obtained in a random selection of size l is al/n . The probability of finding, say all the a CMC molecules in a subset of size l is very very small²⁹ (if $n = 10\,000$, $l = 100$ and $a = 5$ the probability is of the order of 10^{-10}). The first column in the table lists the total number of drug-like molecules in the 10 000 compound dataset. The second column lists (as a percent) the average number of CMC compounds found in all of the 20 library-design experiments conducted using BNN5 with 173 descriptors. We also show within brackets the number of times *all* of the possible CMC compounds are selected in the subset. As is obvious from the table, the increase in the probability of success is very large. Also, as expected, the probability of success decreases as the number of CMC compounds in the set increases to a 100. Also, with 500 drug-like molecules in the original list the probability of finding 100 drug-like molecules in a selection based on BNN5 should be close to 100%. Figure 3 lists 50 compounds from the ACD that appeared in one of the list of top 100 drug-like compounds. This figure provides some clue to the kinds of molecules the network deems to be drug-like from the ACD.

3.6. A Second Experiment in Library Design: Comparison to a Diversity Approach. In the standard diversity approach to selection, a random cutoff point (say 0.3) in say the Tanimoto similarity²¹ is applied, and only those compounds with a similarity co-

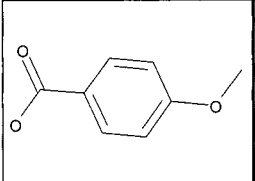
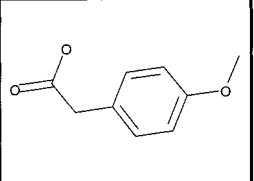
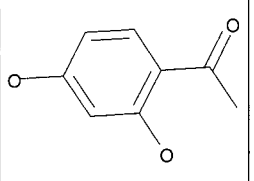
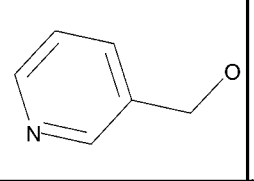
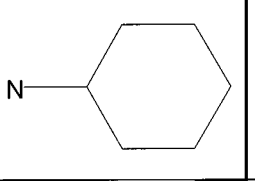
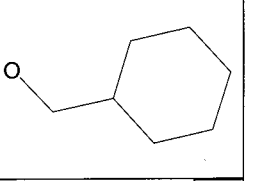
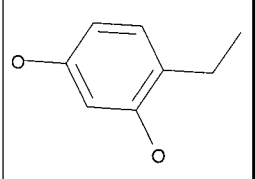
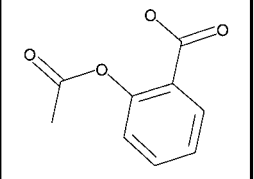
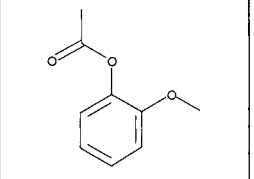
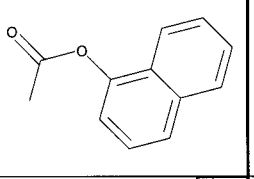
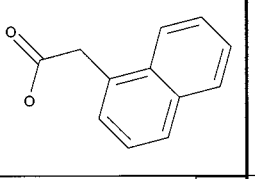
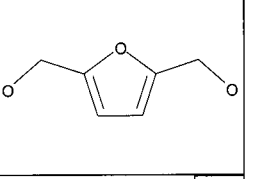
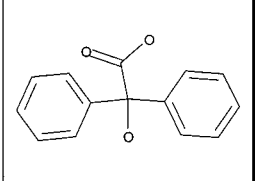
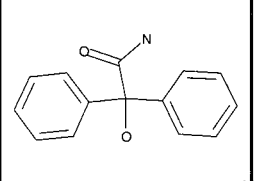
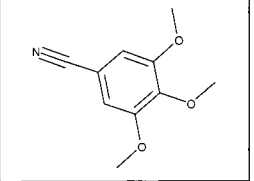
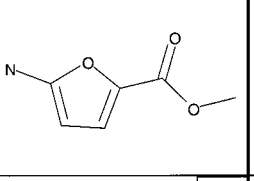
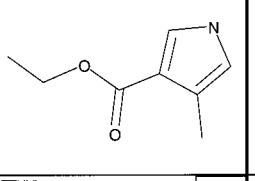
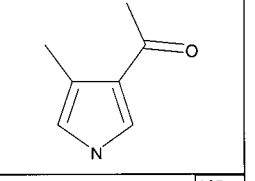
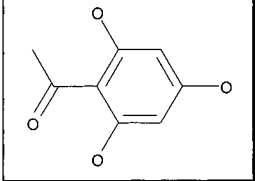
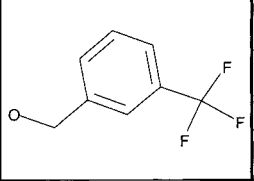
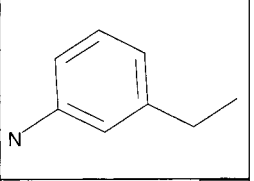
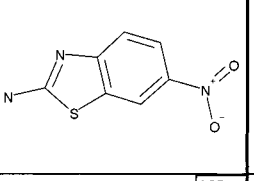
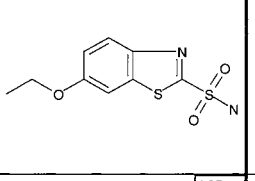
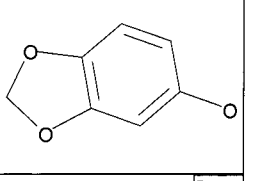
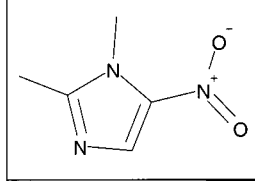
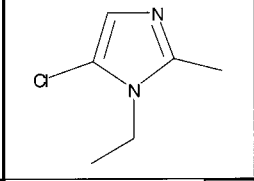
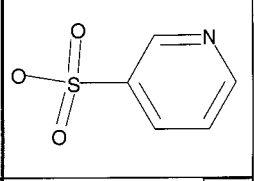
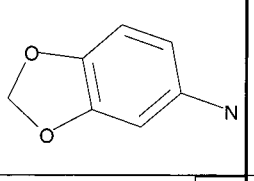
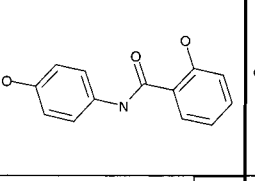
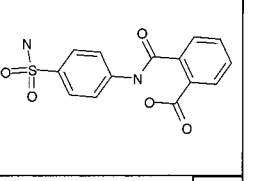
					
P-ANISIC ACID ACD Drug	4-METHOXYPHENYLACETIC ACID ACD Drug	2,4-DIHYDROXYACETOPHENO NE Drug Drug	3-(HYDROXYMETHYL)PYRIDINE Drug Drug	CYCLOHEXYLAMINE ACD ACD	CYCLOHEXYLMETHANOL ACD Drug
					
4-ETHYLRESORCINOL Drug Drug	ACETYSALICYLIC ACID ACD Drug	2-METHOXYPHENYL ACETATE ACD Drug	1-NAPHTHYL ACETATE ACD ACD	1-NAPHTHYLACETIC ACID ACD ACD	2,5-FURANDIMETHANOL Drug Drug
					
BENZILIC ACID Drug Drug	BENZILIC ACID AMIDE Drug ACD	3,4,5-TRIMETHOXYBENZONITRI LE ACD Drug	MAYBRIDGE EN 00155 ACD Drug	ETHYL 4-METHYL-3-PYRROLECARBOX YLATE ACD ACD	MAYBRIDGE BTB 09337 ACD ACD
					
2,4,6-TRIHYDROXY ACETOPHENONE Drug Drug	3-(TRIFLUOROMETHYL)BENZYL ALCOHOL ACD Drug	3-ETHYLANILINE ACD ACD	2-AMINO-6-NITROBENZOTHAZ OLE ACD ACD	6-ETHOXY-2-BENZOTHAZOLES ULFONAMIDE ACD ACD	SESAMOL Drug Drug
					
1,2-DIMETHYL-5-NITROIMIDAZ OLE ACD ACD	5-CHLORO-1-ETHYL-2-METHYL IMIDAZOLE ACD ACD	PYRIDINE-3-SULFONIC ACID Drug Drug	3,4-METHYLENEDIOXYANILINE ACD ACD	N-(4-HYDROXYPHENYL)-SALIC YLAMIDE Drug Drug	4'-SULFAMOYLPHTHALANILIC ACID ACD ACD

Figure 2. Small list of 30 compounds with the consensus predictions from BNN5. Drug-like compounds are labeled "Drug" and nondrug-like compounds "ACD". The first label is based on the 173 descriptor set and the second on the 75 descriptor set.

Table 8. Performance on Designing a Library of Drug-like Molecules of Size 100 from an Original Set of 10 000 Compounds^a

no. of drugs	% times the drugs appear in subset
1	95 (95)
5	81 (35)
10	82 (10)
50	77 (0; max. found 44)
100	72 (0; max. found 79)
500	100 (does not apply)

^a The first column shows the number of compounds from the CMC in the list of 10 000 molecules. The probability of finding at least one active in a random draw of 100 compounds is about 0.01 (if there is one active in the original set of 10 000 compounds). The probability of finding all the drug-like molecules in a random draw gets progressively smaller and gets to be truly miniscule by the time there are 50 CMC molecules in the original set. The second column lists the average number (in percent) of CMC molecules found in 20 different realizations of the 100 compound subsets using BNN5. As an example, if the number of drugs in the original list is 5, then for the 20 different realizations we expect to obtain a maximum of 100 drugs, but we find only 81. The second column also lists, within brackets, the average number (in percent) of times all the molecules are found to be present in the final subset. If this number is zero, we list maximum number of drug-like molecules found in the top 100 drug-like molecules. For the last row it does not make much sense to calculate this number and hence is designated by "does not apply." There is a very substantial improvement over random methods.

efficient lower than the cutoff make it into the selection. One natural question is, how does our method compare with just looking at Tanimoto coefficient similarity? To understand this we compare two sets of compounds, (1) all drug-like compounds in the ACD with the compounds in the CMC and (2) all nondrug-like molecules in the MDDR with the compounds in the CMC. For both sets, we tabulate the Tanimoto coefficient of the most similar CMC compound. For example, for each drug-like compound from the ACD, we determine the Tanimoto coefficient for the most similar molecule from the CMC. These Tanimoto coefficients are shown as a histogram in Figure 4. This figure shows Tanimoto coefficient values based on topological torsions.³⁰ Note that the distribution is peaked about 0.5 and shows a reasonable spread. Clearly, the compounds selected as drug-like by the neural network are not "similar" based on Tanimoto metrics. A very similar histogram is obtained when the analysis is repeated for the nondrug-like molecules in the MDDR and also when calculating Tanimoto similarity based on just the 166 ISIS keys (results not shown). As mentioned before, both sets of descriptors (topological torsions and ISIS keys) are in common use in the diversity literature. This demonstrates that neighborhoods defined by the BNN5 models are not similar to the ones generated by Tanimoto similarity cutoffs based on either topological torsions or the ISIS keys as descriptors. Therefore, new and different information that is not accessible by standard similarity metrics is being generated by our models.

4. Discussion

Some years ago it was shown that only 32 scaffolds describe half of all known drugs.¹⁴ A similar analysis has been carried out for the side chains in known drugs. Again we observe that a small number of moieties account for a large majority of the side chains found in drugs.²⁸ One admittedly speculative interpretation of

these findings is that the "universe" of drug-like molecules may be less diverse and more apprehensible than previously imagined.

In this paper we have introduced and quantified the notion of drug-likeness. Assuming this to be a "property" of a molecule, it can be used to design combinatorial libraries alone or in conjunction with the standard approaches based on diversity. The basic assumption in this work is that the CMC database is a good surrogate for drug-like and the ACD database is a good surrogate for nondrug-like molecules. We have described one of the very few methods that incorporate biological information (we know that CMC molecules show biological activity on a wide range of targets) into computational methods of library design.

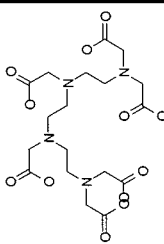
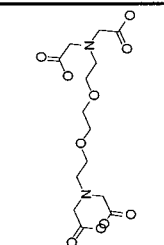
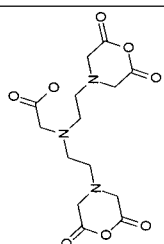
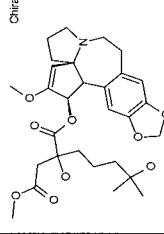
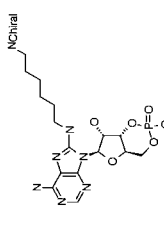
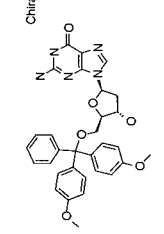
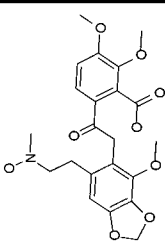
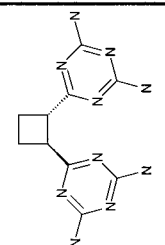
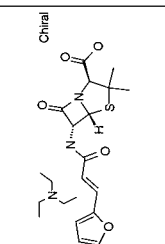
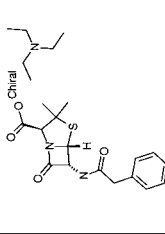
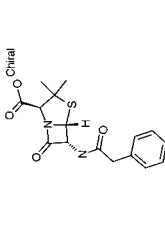
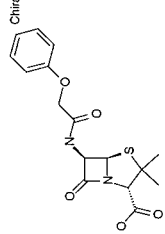
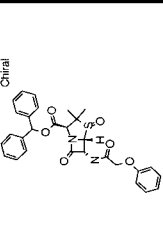
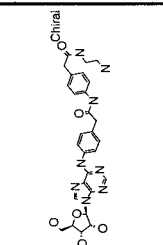
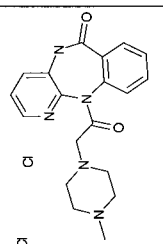
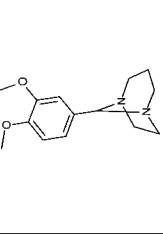
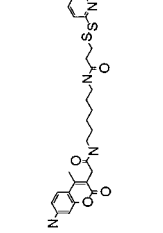
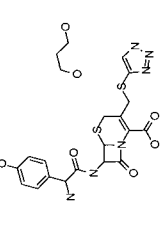
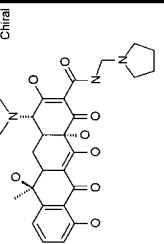
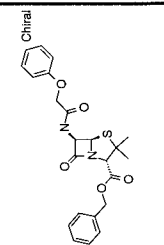
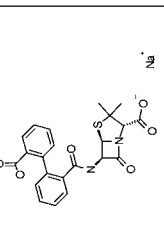
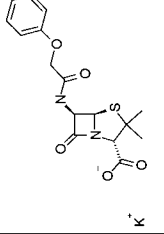
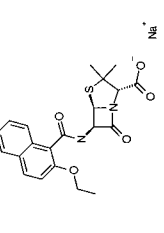
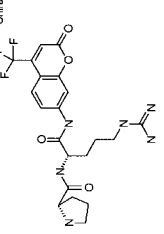
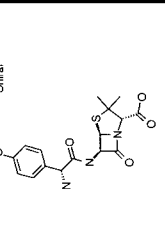
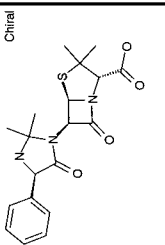
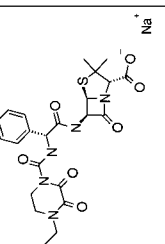
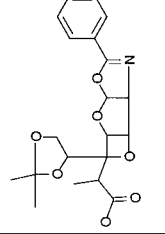
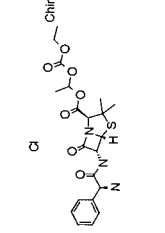
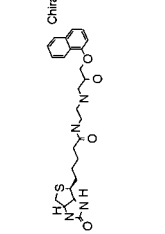
Our results are very encouraging. We have shown that it is possible to cover over 90% of the CMC database using a set of simple 1D/2D descriptors. We have also shown that extrapolations are possible using the models built using BNN5. This is demonstrated by the ability of the models to predict on the MDDR database resulting in about 80% of its contents being deemed drug-like. This is an extremely useful result and provides us reasonable confidence in designing drug-like combinatorial libraries.

Another useful lesson from this work is that we have shown that it is possible to select a set of the most pertinent descriptors. It will be interesting to repeat the library design experiments done by Brown and Martin²¹ using this subset instead of the complete list of 166 ISIS keys that was used in their work.³¹ We have shown that it is possible to build libraries that contain drug-like molecules with anywhere from 3 to 5 orders of magnitude higher probability than random methods. In fact, at worst the networks perform about 72× better than random in selecting drug-like molecules. This can be done while maintaining a reasonable degree of diversity in compound selection based on our results of analyzing Tanimoto coefficients. We have also shown that drug-likeness cannot be understood in terms of closeness in Tanimoto coefficients.

A number of tests indicate that there are real differences between the distributions of compounds in the CMC and the ACD. The results in Figure 2 point out that small context-dependent changes to a molecule can change the classification. It has not been possible to capture in a few words or pictures the characteristics that distinguish drugs from nondrugs. A priori we would not expect such simple, distinguishing characteristics to exist.

Molecules can be represented in a variety of ways. It is reasonable to assume that models based on complimentary sets of descriptors will improve the reliability of predictions. We are, therefore, working on developing a new set of descriptors in order to further enhance the reliability of predictions. Emboldened by this success, we are also working on learning to distinguish between CNS (central nervous system) and non-CNS active compounds and other drug classes.³²

It is possible to argue that we are merely capturing the characteristics of existing drugs and that using this as a filter will be detrimental to exploring new structural motifs. This is a reasonable argument. However, our results do offer some hope. First, it has been

					
TRIETHYLENÉTETRAMINEHEXAACETIC ACID	TRIETHYLENÉSIS(OXYETHYLENÉTRIOLO)TETRAACETIC ACID	DIETHYLENÉTRIAMINOPENTAACETIC DIANHYDRIDE	HOMOHARRINGTONINE	8-(6-AMINOHEXYLAMINO-ADENOSINE 3',5'-CYCLIC MONOPHOSPHATE)	5-(4-(4-DIMETHOXYTRITYL)-2-DEOXYRIBO-FURANOSYL)URACIL
					
N-OXYNORFLORFEN	TRANS-1,2-CYCLOBUTANEDICARBODI-ANAMINE	6-BETA-FURYLACRYLOYLAMIDOPENICILLANIC ACID, TRIETHYLAMMONIUM SALT	TRIETHYL AMMONIUM PENICILLIN	D-BENZYL PENICILLINIC ACID	PHENOXYMETHYLPENICILLINIC ACID
					
PHENOXYMETHYLPENICILLINIC ACID BENZHYDRYL ESTER SULFOXIDE	ADENOSINE AMINE CONGENER	PIREZEPINE DIHYDROCHLORIDE	8-(4-ETHOXY-3-METHOXYPHENYL)-1,5-DIAZABICYCLO(3.2.1)OCTANE	AMCA-HPDP	CEFATRIZINE PROPYLENE GLYCOL
					
ROLITETRACYCLINE	PHENOXYMETHYLPENICILLINIC ACID BENZYL ESTER	2-(2'-CARBOXYPHENYL)BENZOYL-6-AMINOPENICILLANIC ACID SODIUM SALT	PENICILLIN V POTASSIUM	NAFACILLIN SODIUM SALT	DIPEPTIDYL PEPTIDASE I (SUBSTRATE II)
					
AMOXICILLIN	HETACILLIN	PIPERACILLIN SODIUM SALT	MAYBRIDGE NRB 00272	BACAMPICILLIN HYDROCHLORIDE	BIOTIN-PROPIONOLOL ANALOG

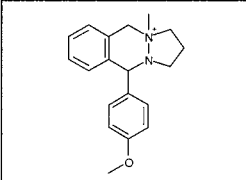
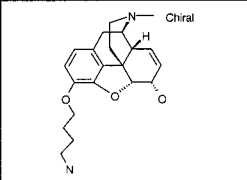
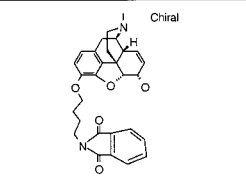
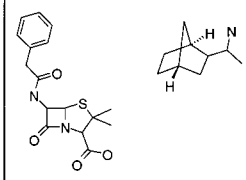
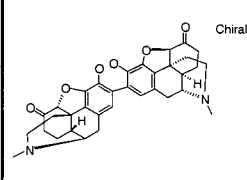
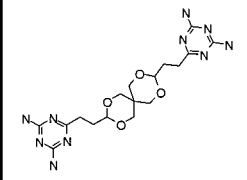
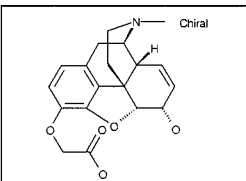
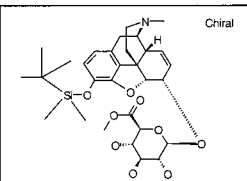
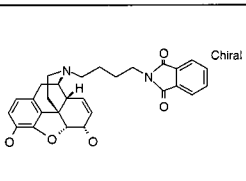
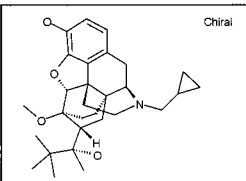
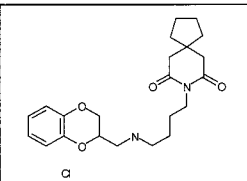
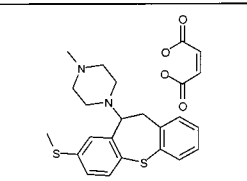
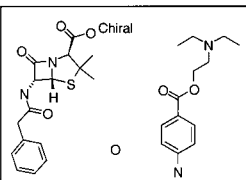
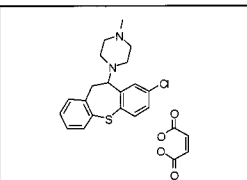
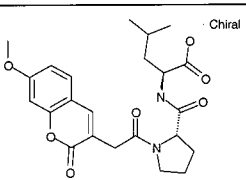
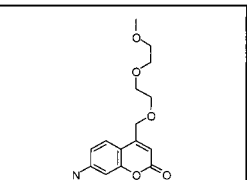
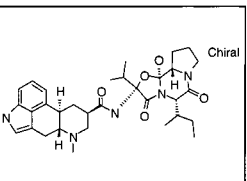
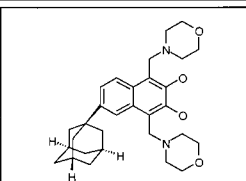
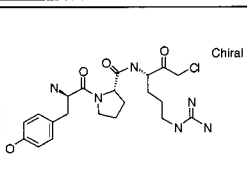
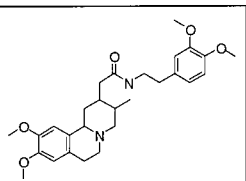
					
(4-METHOXYPHENYL)-METHYL-TETRAHYDRO-1H-PYRAZOLO(1,2-B)PHTHALAZINE-4-IUM IODIDE	3-O-(4-AMINO)BUTYL)MORPHINE	3-O-(4'-PHTHALIMIDO)BUTYL)MORPHINE	DI-ME-OXO-PH-AC-AMINO-THIA-AZA-BICYCLO(3,2,0)HEPTANE-CARBOXYLIC ACID	PSEUDOHYDROMORPHONE	3,9-BIS[2-(3,5-DIAMINO-2,4,6-TRIAZAPHENYL)ETHYL]-2,4,8,10-TETRAOXASPIRO[5,5]UND
					
3-O-CARBOXYMETHYL)MORPHINE	METHYL 3'-O-(T-BUTYLDIMETHYLSILYL)MORPHINE-6-YL	N-(4'-PHTHALIMIDO)BUTYL)NORMORPHINE	BUPRENORPHINE	MDL 72832 HYDROCHLORIDE	METHIOTHEPIN MALEATE
					
PENICILLIN G PROCaine SALT	OCTOCLOTHEPIN MALEATE	MCA-PRO-LEU-OH	7-AMINO-4-(2,5,8-TRIOXANONYL)-COUMARIN	DIHYDROERGOCRYPTINE	
					
SZS TECH L3/298	D-TYR-PRO-ARG-CMK	2-(DI-MEO-3-ME-PYRIDO(2,1-A)ISOQUINOLIN-2-YL)-N-(2-(DIMETHOXY-PH)-ETHYL)-ACETAMIDE			

Figure 3. List of 50 compounds from the ACD that were classified with high probability of being drug-like. See the text for more details.

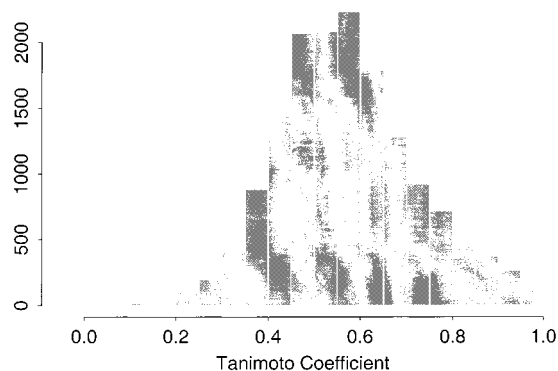


Figure 4. Histogram of Tanimoto coefficients based on topological torsions of the most similar CMC molecule for each of the drug-like molecules from the ACD. This demonstrates that there is no simple relationship between drug-likeness and standard 2D similarity measures of molecules.

possible for us to extrapolate with good success to the MDDR database after learning on the CMC. This shows that the models we have built are capable of recognizing a broad class of structural motifs. Second, the method advocated in this paper can work well in conjunction with standard diversity protocols, especially by considering lower probability drug-like molecules. Third, from our results on exploring similarity using standard methods it is evident that our method does not behave in a similar fashion to standard notions of similarity used in diversity calculations. Finally, one could always choose to select say a portion of the compounds using this filter while reserving the rest for other methods.

Acknowledgment. One of us (Ajay) would like to thank Guy Bemis and David Rogers (of MSI) for interesting discussions on the subject of this paper. We would also like to thank Paul Charifson for useful comments on the manuscript.

Supporting Information Available: Detailed description of C4.5 and Bayesian neural networks (13 pages). Ordering information is given on any current masthead page.

References

- Gordon, E. M. Libraries of non-polymeric organic molecules. *Curr. Opin. Biotechnol.* **1995**, *6*, 624–31.
- Dolle, R. E. Discovery of enzyme inhibitors through combinatorial chemistry. *Mol. Diversity* **1997**, *2*, 223–36.
- Brown, D. Future pathways for combinatorial chemistry. *Mol. Diversity* **1997**, *2*, 217–222.
- Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431–1436.
- Matter, H. Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219–1229.
- Lewis, R. A.; Mason, J. S.; McLay, I. M. Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (dpd) approach. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 599–614.
- Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750–763.
- Murcko, M. A. Recent advances in ligand design methods. In *Reviews in Computational Chemistry, Vol. 11*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley: New York, in press.
- Bristol, D. W.; Wachsmann, J. T.; Greenwell, A. The NIEHS predictive-toxicology evaluation project. *Environ. Health Perspect.* **1996**, *104*, 1001–10.
- Ridings, J. E.; Manallack, D. T.; Saunders, M. R.; Baldwin, J. A.; Livingstone, D. J. Multivariate quantitative structure-toxicity relationships in a series of dopamine mimetics. *Toxicology* **1992**, *76*, 209–17.
- Comprehensive Medicinal Chemistry Release 94. 1 is available from MDL Information Systems Inc., San Leandro, CA 94577. An electronic database of volume 6 of *Comprehensive Medicinal Chemistry* published by Pergamon Press in March 1990 contains drugs already in the market.
- MACCS-II Drug Data Report is available from MDL Information Systems Inc., San Leandro, CA 94577. An electronic database version of the prous science publishers journal *Drug Data Report*, extracted from issues starting mid-1988, contains biologically active compounds in the early stages of drug development.
- Available Chemicals Directory is available from MDL Information Systems Inc., San Leandro, CA, and contains specialty and bulk chemicals from commercial sources.
- Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- 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.
- Hertz, J.; Krogh, A.; Palmer, R. G. *Introduction to the Theory of Neural Computation*; Addison Wesley: Redwood City, CA, 1991.
- Quinlan, J. R. *CA. 5: Programs for machine learning*; Morgan Kaufmann: San Mateo, CA, 1993.
- Gillet, V.; Willett, P.; Bradshaw, J. Development of bioactivity profiles for use in compound selection. 11th ACS National Meeting New Orleans, LA. March 24–28, 1996.
- Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley: New York, 1991; Vol. 2, pp 367–422.
- SSKEYS, MDL Information Systems Inc., San Leandro, CA.
- Brown, R.; Martin, Y. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.
- Neal, R. An improved acceptance procedure for the hybrid monte-carlo algorithm. *J. Comput. Phys.* **1994**, *111*, 194–203.
- Buntine, W. L.; Weigand, A. S. Bayesian back-propagation. *Complex Systems* **1991**, *5*, 603–43.
- Tierney, L. Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.* **1994**, *22*, 1701–62.
- Cleveland, W. S. *Visualizing data*; Hobart Press: Summit, NJ, 1993.
- This pruning exercise does not significantly deplete the databases. For example, after removing the “nondrugs” from the CMC, there are about 4700 compounds in the CMC for which we could calculate the log *p* values. About 150 compounds from this had an MW > 600, 200 with both MW > 600 and $-4.0 \geq \log P \geq 11.0$, and 203 compounds with MW > 600, $-4.0 \geq \log P \geq 11.0$, and $^2\kappa_{\alpha} > 30$.
- Johnson, R. A.; Wichern, D. *Applied Multivariate Statistical Analysis*, 3rd ed.; Prentice-Hall: Englewood Cliffs, NJ, 1992.
- Bemis, G. W.; Murcko, M. A. The properties of known drugs. 2. sidechains. Submitted.
- If there are *n* compounds in total, with *a* actives, and we select a library of size *l*, then the probability of finding all *a* in a random draw (without replacement) of size *l* is given by $[(n-a)!/n!(l-a)!]$.
- Nilakantan, R.; Bauman, N.; Dixon, J.; Venkataraghavan, R. Topological Torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
- Ajay. Linking activity-space and descriptor-space: An exploration. Work in progress.
- Ajay; Murcko, M. Designing combinatorial libraries with cns activity. Submitted.
- Risannen, J. *Stochastic Complexity in Statistical Enquiry*; World Scientific, 1989.
- Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*, Cambridge University Press: Cambridge, 1988.
- MacKay, D. J. C. Bayesian interpolation. *Neural Comput.* **1992**, *4*, 415–447.
- MacKay, D. J. C. A practical Bayesian framework for back-propagation networks. *Neural Comput.* **1992**, *4*, 448–472.
- MacKay, D. J. C. Probable networks and plausible predictions—a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* **1995**, *6*, 469–505.
- Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087–92.
- Duane, S.; Kennedy, A. D.; Pendleton, B. J.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216–22.