Differential Shannon Entropy Analysis Identifies Molecular Property Descriptors that Predict Aqueous Solubility of Synthetic Compounds with High Accuracy in Binary QSAR Calculations

Florence L. Stahura,† Jeffrey W. Godden,† and Jürgen Bajorath*,#

Computer-Aided Drug Discovery, Albany Molecular Bothell Research Center, 18804 North Creek Parkway, Bothell, Washington 98011, and Albany Molecular Bothell Research Center and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received November 1, 2001

Prediction of aqueous solubility of organic molecules by binary QSAR was used as a test case for a recently introduced entropy-based descriptor selection method. Property descriptors suitable for solubility predictions were exclusively selected on the basis of Shannon entropy calculations in molecular learning sets, not taking any other information into account. Sets of only five or 10 2D descriptors with largest entropy differences between molecules above or below a defined solubility threshold yielded consistently high prediction accuracy between 80% and 90% in binary QSAR calculations, regardless of the threshold values applied. The top five descriptors with largest differential Shannon entropy (DSE) values achieved an average prediction accuracy of 88%. These findings suggest that differences in entropy and relative information content of descriptors in compared compound data sets correlate with significant differences in physical properties and support the practical relevance of entropy-based descriptor selection routines. The study also demonstrates that binary QSAR methodology can be effectively used to classify small molecules according to aqueous solubility.

INTRODUCTION

Many tasks in chemoinformatics research depend on the use of chemical descriptors to capture and represent molecular structure and properties.[1,2] These include, among others, the definition of abstract chemical spaces for diversity analysis, clustering and classification of compounds, or design of computational tools for the assessment of molecular similarity and virtual screening.[2,3] In this context, an important topic is how to rationalize the selection of descriptors for specific computational applications beyond chemical intuition.[3] In studies designed to develop a quantitative computational metric for descriptor selection, the Shannon entropy (SE) concept,[4] originally introduced in digital communication theory, has recently been adapted[5] and extended.[6,7] SE calculations reduce compound database distributions of molecular descriptors to their information content.[5] Subsequent calculation of differential Shannon entropy (DSE) values makes it possible to quantitatively compare information content and value range distributions of descriptors in different compound data sets, regardless of their units and values.[6] For database comparisons, SE-DSE schemes can be established to classify descriptors according to their relative information content.[7] For example, descriptors belonging to the "high-high" SE-DSE category have consistently high information content in databases under comparison and are sensitive to systematic differences between these data sets (for example, chemical reagents versus drug-like molecules).[7] Thus, this classification scheme is thought to permit the identification of descriptors that are suitable to, for example, differentiate between compounds belonging to distinct chemical classes.

In evaluating the SE-DSE concept, an important question has been whether differences between descriptor settings that are detected at the level of entropy calculations indeed correlate with physically significant changes in molecular properties. In other words, can the entropic formalism be used to identify descriptors that respond to specific physical differences between molecules? Are differences in relative information content are of practical relevance? In an initial study to address these questions, we have selected descriptors with favorable SE characteristics based on an analysis of natural product and synthetic compound databases.[8] Combinations of these descriptors could be used to systematically distinguish between molecules from natural and synthetic sources.[8] These findings provided initial evidence for the applicability of entropy-based descriptor selection schemes. However, classification of natural and synthetic molecules can also be accomplished by utilizing differences in the frequency of occurrence of substructural fragments,[8,9] and, therefore, these predictions do not depend on the identification of suitable molecular property descriptors. Thus, to address a conceptually different and perhaps more relevant problem, we have now gone a step further and investigated descriptor entropy analysis in the context of aqueous solubility predictions of synthetic compounds, for several reasons. Aqueous solubility is a good example of a "physicochemical" phenomenon that can be addressed using molecular property descriptors. In addition, since aqueous solubility is relatively easy to measure experimentally, high quality data sets can be assembled for derivation and evaluation of predictive models. Moreover, solubility problems are being recognized

* Corresponding author phone: (425)424-7297; fax: (425)424-7299; e-mail: jurgen.bajorath@albmolecular.com.

† Computer-Aided Drug Discovery, Albany Molecular Bothell Research Center.

Albany Molecular Bothell Research Center and University of Washington.

10.1021/ci010243q CCC: $22.00
Published on Web 02/19/2002

as major bottlenecks in lead identification and optimization,[10,11] and, in consequence, prediction of aqueous solubility of small molecules is currently the focal point of many computational studies.[11]

Contemporary computational approaches to the prediction of aqueous solubility can essentially be divided intro three different categories, QSPR or linear regression-type models,[12−15] neural network simulations,[15−18] and molecular fragment-based or group contribution methods.[19,20] In the latter case, solubility values are assigned to structural fragments, their occurrence in molecules is determined, and aqueous solubility is estimated incrementally. In contrast to QSPR and neural network approaches, group contribution methods do not depend on the use of physical property descriptors. For QSPR and neural network calculations, initial descriptor sets have often been chosen intuitively (e.g., those accounting for polar or hydrophobic character, charge, solvent-accessible surface area, hydrogen bonding, logP(o/w) etc.) or selected to best fit multiple linear regression or neural network models for learning sets.[15] This "reductive" approach typically involves an initial evaluation of large numbers (i.e., several hundred) of 2D or 3D molecular descriptors and the identification of those descriptors that best account for solubility values of specific molecular data sets. Molecular simulations of test molecules in water have also been employed to identify suitable descriptors.[12] Among the best-performing descriptors identified in various studies were those combining electronic and topological molecular information,[17,18] in particular, Hall and Kier E-state descriptors.[21]

In this study, we have analyzed aqueous solubility of synthetic and drug-like molecules as a test case for entropy-based descriptor selection. Our key question has been whether entropy analysis could identify descriptors to accurately predict a physical property. Applying different solubility threshold values, we have compared entropic characteristics of more than 100 2D molecular descriptors in data sets consisting of "soluble" and "insoluble" molecules. Descriptors were selected exclusively on the basis of significant DSE values, not taking any intuitive criteria into account, and then used to compute binary QSAR[22] models for the prediction of molecules in test sets. Binary QSAR was selected as a prediction method for two reasons. This approach is conventionally used to differentiate between "active" and "inactive" molecules but not to predict physical properties. Thus, solubility prediction represents a novel type of application. In addition, the method can be conveniently applied to test many different descriptors and perform predictions for large numbers of molecules. In our entropy analysis, we found that suitable descriptors varied in part with selected solubility threshold values and changes in the composition of molecular learning sets. In our binary QSAR calculations, regardless of the solubility threshold values applied, >80% prediction accuracy has consistently been achieved, with only five or 10 descriptors selected on the basis of their DSE values. Thus, the analysis provides substantial support for the relevance of the SE-DSE concept for descriptor selection. The calculations reported herein should also add to the spectrum of computational methods for prediction of aqueous solubility, since binary QSAR methodology has, as mentioned above, not yet been evaluated for this purpose. While these "nonclassical" QSAR models do not predict exact solubility values, they offer variable classification schemes. Since their predictive value is high, these models can be applied in practice, for example, to calculate solubility profiles for large compound libraries or to identify compounds that are predicted to have solubility problems at certain concentrations. Other important aspects of the obtained results are that only a few 2D and implicit 3D molecular descriptors were sufficient to yield accurate predictions and that 3D structures or models of test molecules were not required to achieve a high level of accuracy.

MATERIALS AND METHODS

SE-DSE Concept in Descriptor Analysis. Our implementation of the Shannon entropy concept[4] for descriptor and database comparison and SE and DSE calculations has been described in detail.[5,6] SE analysis reduces a given data distribution to its information content, regardless of data units or value ranges. For example, molecular descriptors can be calculated in large compound databases, their value distributions can be graphically represented, and then transformed into information content. By doing so, the variability of molecular descriptors in different databases can be compared. Shannon entropy is defined as

$$SE = -\sum p_i \log_2 p_i$$

with p being the probability of a data point to adopt a value within a specific data interval i. In this formulation, $\log_2$ is a binary scale factor of probability (i.e., how many yes/no questions do we need to ask to identify the data interval into which each count falls?), allowing SE to be considered as a metric of digital information content. The probability p is calculated as

$$p_i = c_i / \sum c_i$$

where c_i is the count of data occurrences for bin i.

SE values can be calculated for any data set presented in evenly spaced data intervals (bins) and compared, provided the binning scheme is uniform. Therefore, histograms where each data range is divided into the same number of bins provide a particularly convenient form of data representation for SE calculations.[5] In histogram representations, largest possible SE (maximum information content) is obtained when data points are evenly distributed over all data intervals, and maximum SE thus corresponds to $\log_2$ of the number of histogram bins. It follows that bin-independent or scaled SE values (between 0 and 1) are obtained by dividing SE by $\log_2$ of the number of bins.[7] SE values for different data sets (for example, those of a molecular descriptor calculated for two distinct compound databases) can best be compared by calculation of Differential Shannon entropy,[6] defined as

$$\mathrm{DSE} = \mathrm{SE}_{AB} - (\mathrm{SE}_A + \mathrm{SE}_B)/2$$

In this formulation, SE_A and SE_B are SE values for two sets of compounds (A and B), and SE_{AB} represents SE calculated for the combined population (from a single histogram). While database-specific descriptor variability directly correlates with SE values, calculation of DSE captures differences in value range distributions in compared databases. For descriptor analysis, it can be rationalized as an increase or decrease in descriptor variability when two populations are combined. If data distributions are comple-

mentary, large DSE values will be obtained, reflecting significant differences in descriptor entropy and value distribution in the compared data sets. By contrast, if the combined distributions do not diverge from a single population, DSE becomes zero. To summarize, a key aspect of DSE analysis is that entropy values and thus information content depend on the variability of descriptor distributions and the way they are monitored. For example, representing the database values of a specific descriptor on a linear or logarithmic scale will change histogram distributions and thus their SE values. Regardless, DSE analysis is designed to detect significant differences in databases distributions of molecular descriptor.

In this study, we have used different aqueous solubility threshold values to divide a learning set into two subsets, "soluble" and "insoluble" compounds, consistent with the principles of binary QSAR calculations, as discussed below. Therefore, SE_A and SE_B correspond to SE_S for "soluble" and SE_{INS} for "insoluble" compounds, respectively. For each threshold value and corresponding learning set (with varying numerical composition; see below), molecular descriptors were calculated, and their value distributions were captured in histograms having consistently 25 bins. On the basis of this data representation, SE and DSE values were determined for each descriptor. For each solubility threshold value, descriptors were ranked according to largest DSE values, which correspond to most significant differences between soluble and insoluble compounds.

Compound Data Sets and Solubility Threshold Values. A total of 650 molecules with known aqueous solubility, expressed as logS, where S is the solubility in mol/L, were assembled from the literature,[13−17] including synthetic and drug-like molecules. Care was taken to only include compounds for which solubility was experimentally measured at a constant temperature (25 ± 1 °C). To ensure high quality of the data set, the temperature and also the -logS value reported in the literature were confirmed for each selected compound in the PHYSPROP database.[23] The 650 compounds were randomly divided into a training set of 550 and a test set of 100 molecules. A histogram analysis confirmed that the solubility ranges covered in the training and test sets were very similar, i.e., solubility values (in -log S units) ranged from −0.52 to 10.8 for the training and from −0.96 to 9.47 for the test set. Both training and test sets were divided into soluble and insoluble molecules using five solubility threshold values, 1 mM (i.e. -logS = 3), 5 mM (i.e. -logS = 2.3), 10 mM (i.e. -logS = 2), 50 mM (i.e. -logS = 1.3), and 100 mM (i.e. -logS = 1). These threshold values were chosen to capture a solubility range into which many known drugs fall[23] and, furthermore, because compounds not soluble at 10 mM often cause problems with reproducibility of screening assays.[10] Depending on the threshold value, the numerical compositions of learning and prediction sets change, as summarized in Table 1. With increasing threshold value, the number of compounds classified as soluble decreases. For both the 1 mM and 5 mM thresholds, the number of soluble and insoluble compounds is very similar. In this case, when assembling training sets for binary QSAR analysis, the most important requirement was that the threshold value was well within the distribution of solubility values of selected compounds. By contrast, absolute solubility values of compounds were less important.

Table 1. Numerical Composition of Training and Test Sets at Varying Solubility Threshold Values[a]

solubility threshold (mM)	no. of compounds "soluble"	no. of compounds "insoluble"
1	283	267
	54	***46***
5	222	328
	44	***56***
10	186	364
	34	***66***
50	126	424
	16	***84***
100	96	454
	13	***87***

[a] The numbers of compounds in the soluble and insoluble subsets of the test set (100 molecules) for each threshold value are reported in bold italics. Corresponding numbers are provided for the learning set (550 molecules).

Molecular Descriptors. A total of 148 molecular descriptors were subjected to DSE analysis of the learning set. This descriptor set was described previously[5] and contains diverse types of property descriptors including various connectivity, shape, and molecular graph indices, charge, hydrogen bond, and implicit surface and topology descriptors as well as descriptors accounting for polar, aromatic, or hydrophobic character. Table 2 defines a subset of these descriptors that are discussed in this study based on their DSE characteristics. All descriptors included in this set can be calculated from 2D representations of molecules. A recent addition to our basic descriptor set has been a new class of implicit 3D descriptors that map properties on molecular surface areas approximated from 2D structures,[24] termed here "Labute descriptors". Electrotopological or E-state descriptors,[21] shown to be powerful in predicting solubility in other investigations, as discussed above, were not available for this study. We decided to initially omit explicit 3D descriptors from our calculations, to reduce the magnitude of descriptor comparison and selection, and include 3D descriptors later on if no satisfactory results could be obtained. However, as demonstrated in the following, this was not the case, and, therefore, an extension of our 2D-focused descriptor set was not required.

Binary QSAR Models. The binary QSAR methodology and applications have been described in a number of recent publications.[8,22,25−27] Briefly, based on Bayes' Theorem,[28] binary QSAR correlates structural features and properties of molecules, captured by descriptor combinations, with a probability to adopt a state within a binary classification scheme,[22] most commonly an "active" or "inactive" state (i.e., QSAR-like correlation of structure and biological activity). Based on a training set, binary QSAR calculates a probability density of compounds to be either active or inactive. This is done with the aid of principal component analysis (PCA)[29] of molecular descriptor space that decorrelates and normalizes a set of descriptors which then leads to a probability density function. Each descriptor combination submitted to binary QSAR analysis defines a probability function that is then used as a model to predict the state of test compounds.

For the purpose of our analysis, the binary molecular states were defined as "soluble" (= 1) and "insoluble" (= 0), rather than "active" or "inactive", and we used combinations of descriptors identified by DSE analysis to generate models

Table 2. Molecular Descriptors with Largest Average DSE Values[a]

av DSE	descriptor	definition
0.568	SlogP	log octanol/water partition coefficient
0.554	a_hyd	number of hydrophobic atoms
0.542	logP(o/w)	alternative log octanol/water partition coefficient
0.542	PEOE_VSA_NEG	total negatively charged vdW surface area
0.526	PEOE_VSA-1	vdW surface area with atomic partial charge $-0.1 \le q < -0.05$
0.494	SMR	molar refractivity
0.492	chi1v	atomic valence connectivity index (order 1)
0.492	vsa_hyd	hydrophobic vdW surface area
0.482	mr	alternative formulation of molar refractivity
0.472	chi0v	atomic valence connectivity index (order 0)
0.470	PEOE_VSA_HYD	total hydrophobic vdW surface area
0.464	chi1_C	carbon connectivity index (order 1)
0.454	weinerPath	Weiner path number
0.450	vdw_vol	vdW volume
0.448	chi1v_C	carbon valence connectivity index (order 1)
0.446	a_nC	number of carbon atoms
0.442	apol	sum of atomic polarizabilities
0.440	zagreb	Zagreb index
0.430	chi0_C	carbon connectivity index (order 0)
0.426	SlogP_VSA7	vdW surface with $0.25 < \text{SlogP} \le 0.30$
0.422	b_heavy	number of bonds between heavy atoms
0.422	weinerPol	Weiner polarity number
0.416	Weight	molecular weight
0.404	chi1	atomic connectivity index (order 1)
0.384	chi0v_C	carbon valence connectivity index (order 0)
0.382	SMR_VSA5	vdW surface area with $0.44 < \text{SlogP} \le 0.485$
0.376	a_heavy	number of heavy atoms
0.374	Q_VSA_HYD	total hydrophobic vdW surface area
0.374	Q_VSA_POS	total positive vdW surface area
0.374	vdw_area	vdW surface area (A**2)
0.364	chi0	atomic connectivity index (order 0)
0.348	b_count	number of bonds
0.342	a_aro	number of aromatic atoms
0.340	b_ar	number of aromatic bonds
0.332	SlogP_VSA6	vdW surface area with $0.20 < \text{SlogP} \le 0.25$

[a] DSE values were averaged over all five solubility threshold values. "_VSA_" indicates "Labute descriptors",[24] and "vdW" stands for "van der Waals".

from our learning sets and predict the solubility of test compounds. Since the probability function produces continuous values between 0 and 1, a cutoff-value of 0.5 was used to discriminate between soluble (> 0.5) and insoluble (<0.5) molecules. In our calculations, we did not limit the number of principal components derived from each descriptor combination and applied a smoothing factor of 0.08 to each obtained probability function.[22] For each solubility threshold value, six binary QSAR models were built with 5, 10, 15, 20, 25, or 30 descriptors, respectively, starting with the top five descriptors having largest DSE values. The 10 descriptor models then used the top five descriptors and descriptors ranked six to ten, the 15 descriptor models the top 10 descriptors and the next five, and so on. Each model was used to predict the solubility classification of the test set according to the different threshold values. To evaluate the performance of the different models, prediction accuracies were calculated as follows:

$$\text{PA} = \text{overall prediction accuracy} = (\text{S} + \text{INS})/\text{NT}$$

$$\text{PA}_\text{S} = \text{PA for soluble compounds} = \text{S}/\text{NT}_\text{S}$$

$$\text{PA}_\text{INS} = \text{PA for insoluble compounds} = \text{INS}/\text{NT}_\text{INS}$$

"S" is the number of correctly identified soluble molecules, and "INS" is the number of correctly identified insoluble molecules. "NT_S" is the total number of soluble molecules, "NT_INS" is the total number of insoluble molecules, and "NT" is the total number of compounds (i.e., 100 for the test set).

Calculations. Descriptor values for all database compounds and binary QSAR models were calculated using the Molecular Operating Environment (MOE).[30] Entropy analysis of descriptor distributions was performed with programs written by the authors.

RESULTS AND DISCUSSION

The Concept of Entropy-Based Descriptor Selection. With the SE-DSE concept, an entropic formalism has been introduced as a metric for selection of property descriptors, based on the analysis of the distribution of their values in compound databases. Are there related entropy-based concepts? In early studies, Willett and colleagues made use of "relative entropy"[31] calculations to determine angle and distance ranges or substructural fragments that occurred with similar frequency in compound databases and could be used as screens in 3D database searching.[32,33] From a statistical point of view, our Shannon entropy implementation is probably most similar to the Kullback-Leibler function[34] that determines the similarity of a statistical model and a true data distribution. However, this function is not suitable for the development of descriptor selection algorithms because its value depends on which of the distributions is considered the reference, and, in addition, it is not defined if the model

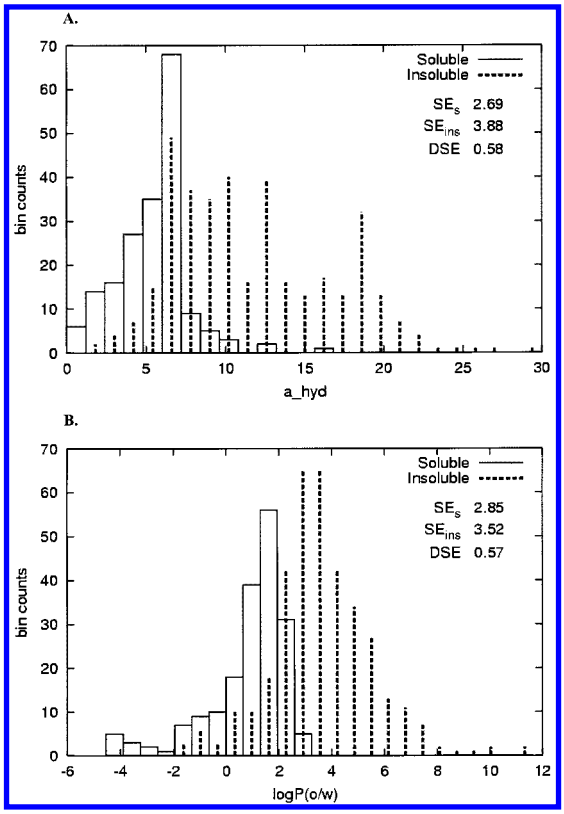

Figure 1. Histograms of descriptor distributions. Two representative examples are shown for descriptors having large DSE values for comparison of learning set compounds with aqueous solubility above (soluble) or below (insoluble) a 10 mM threshold value. Descriptor "a_hyd" (A) counts the number of hydrophobic atoms in a molecule (which is, on average, larger in the insoluble set), and "logP (o/w)" (B) is the well-known octanol/water partition coefficient (for descriptor definitions, see also Table 2), the values of which are also larger for the insoluble set, as one would expect. These two descriptors are among the most intuitive examples with significant differences between the compared compound sets, as revealed by SE-DSE calculations.

distribution has zero probability in a given data interval, which is frequently seen in descriptor analysis. What makes the SE-DSE approach unique in this context is its ability to quantitatively compare the information content of descriptors, even if their units and value ranges are very different.

The SE-DSE selection concept is based on the premise that descriptors with differences in relative information content and value range distributions in compared compound collections are sensitive to systematic chemical differences between these compound sets. Thus, descriptors having these features are most likely to discriminate between different compound classes. The relevance of this concept and its practical applicability to predict physicochemical characteristics of test compounds have been focal points of this study. Thus, we have calculated and compared many descriptors in five solubility threshold-dependent learning sets. In each case, SE values were calculated separately for the soluble and insoluble compound subset, DSE values were determined, and descriptors were ranked according to largest DSE values.

Descriptor Distributions, DSE Values, and Preferred Descriptors. Figure 1 shows representative examples of descriptor distributions that yielded largest DSE values in our subset comparisons. Based on a statistical analysis of DSE value distributions for descriptor analysis in databases containing several hundred thousand compounds, DSE values of ~0.2 or greater belong to the "high DSE" category,[7] and we observe here many values in the 0.3−0.5 DSE range. While the number of molecules that can be compared using the SE-DSE metric is essentially unlimited, it should be noted that, in this case, clear trends were obtained by comparing a total of only 550 compounds with known solubility in our learning sets. Table 2 lists 35 descriptors with largest DSE values when their averages were calculated over all five solubility thresholds. On the basis of our analysis, these descriptors are consistently most responsive to systematic differences between compounds classified as soluble or insoluble according to varying solubility thresholds.

Intuition versus DSE Selection. As one would expect, a number of these descriptors, but certainly not all of them, capture hydrophilic or hydrophobic character in various ways. In some cases, preferred descriptors are intuitive, for example, they directly account for hydrophobic character or logP, which is well in accord with commonly accepted physicochemical models of aqueous solubility. However, in other cases, for example, connectivity indices, molecular graphs, or van der Waals volume, it is not immediately evident why these descriptors should be important to distinguish between different solubility levels. Therefore, these descriptors would have been very difficult to select a priori, based on mechanistic considerations, and without these calculations. This illustrates the importance of "unbiased" descriptor selection beyond intuition. On the basis of DSE analysis, a number of in part very different but well performing descriptors could be selected, the contributions of at least some of which cannot be rationalized in simple chemical terms. It is worth noting that eight Labute descriptors[24] appear in the averaged top 35 list that approximate various surface properties and are fairly complex in their design. Thus, this class of descriptors is quite sensitive to solubility-related differences in physicochemical properties.

Descriptor Ranking at Varying Solubility Threshold Levels. Although approximately 20 of the 35 descriptors listed in Table 2 are also always present in the top 30 list for each solubility threshold value, their ranking and relative importance change significantly depending on the threshold. This is shown in Table 3, which reports the top 10 most important descriptors at each of the five threshold levels. For example, only a few descriptors (e.g., PEOE_VSA_NEG, SMR) occur in all top 10 lists, and logP(o/w) or a_hyd, which are among the most variable and thus discriminatory descriptors at low solubility thresholds, disappear from the top 10 lists at the 50 and/or 100 mM solubility threshold values. The comparison suggests that DSE analysis is sensitive to subtle differences in molecular properties between the studied compound sets. In fact, the subsets of soluble and insoluble compounds compared in this study are not completely distinct but overlapping.

The entropy values reported in Table 3 also reveal another trend: with increasing solubility threshold values, many DSE values also increase. For example, DSE values for descriptor PEOE_VSA_NEG, which occurs in all five top 10 lists, steadily increase from 0.39 to 0.65 over the solubility threshold range. For this descriptor and others, SE values for the soluble subsets decrease with increasing threshold values, whereas SE values for the insoluble subset essentially

Table 3. Most Variable Descriptors at Varying Solubility Threshold Values[a]

A. 1 mM				B. 5 mM				C. 10 mM				D. 50 mM				E. 100 mM			
descriptors	DSE	SE_S	SE_{INS}	descriptors	DSE	SE_S	SE_{INS}	descriptors	DSE	SE_S	SE_{INS}	descriptors	DSE	SE_S	SE_{INS}	descriptors	DSE	SE_S	SE_{INS}
logP(o/w)	0.57	3.0	3.4	a_hyd	0.54	2.8	3.9	PEOE_VSA_NEG	0.58	2.3	3.5	PEOE_VSA-1	0.64	2.2	3.8	SlogP_VSA7	0.78	1.6	3.6
SlogP	0.57	2.7	3.2	logP(o/w)	0.53	3.0	3.5	a_hyd	0.58	2.7	3.9	PEOE_VSA_NEG	0.63	2.2	3.5	PEOE_VSA-1	0.77	2.0	3.8
vsa_hyd	0.49	3.0	3.5	SlogP	0.51	2.7	3.4	logP(o/w)	0.57	2.8	3.5	SlogP_VSA7	0.61	1.9	3.6	weinerPath	0.67	0.50	2.2
a_hyd	0.45	3.0	3.8	vsa_hyd	0.48	3.0	3.6	SlogP	0.55	2.6	3.4	a_hyd	0.61	2.7	3.8	weinerPol	0.67	2.4	4.0
PEOE_VSA_HYD	0.44	3.0	3.5	PEOE_VSA_NEG	0.46	2.6	3.5	PEOE_VSA-1	0.53	2.4	3.8	SlogP	0.58	2.4	3.4	Zagreb	0.66	2.7	4.1
PEOE_VSA_NEG	0.39	2.8	3.4	PEOE_VSA_HYD	0.45	3.0	3.6	chi1v	0.51	2.7	3.9	Zagreb	0.58	2.8	4.1	SMR	0.66	2.7	3.9
chi1v_C	0.34	2.9	3.6	chi1v	0.43	2.8	3.9	vsa_hyd	0.50	2.9	3.6	chi1v	0.57	2.6	3.8	PEOE_VSA_NEG	0.65	2.2	3.5
chi1v	0.34	3.0	3.9	chi1v_C	0.42	2.7	3.6	PEOE_VSA_HYD	0.48	2.8	3.6	weinerPath	0.57	0.6	2.3	b_heavy	0.64	2.8	4.1
SMR	0.34	3.2	4.0	SMR	0.42	3.0	4.0	chi1v_C	0.48	2.6	3.6	SMR	0.56	2.8	4.0	mr	0.63	2.7	3.9
mr	0.33	3.2	4.0	chi1_C	0.41	3.0	3.9	SMR	0.48	2.9	4.0	weinerPol	0.56	2.6	4.0	chi0v	0.62	2.6	3.8

[a] For each solubility threshold value, the 10 descriptors with largest DSE values are listed, obtained by comparison of the respective soluble and insoluble subsets of our training set (see also Table 1).

remain constant. These effects are a consequence of the increasingly skewed numerical composition of the soluble and insoluble subsets when solubility threshold values increase, as documented in Table 1. At low threshold values of 1 mM and 5 mM, these subsets are almost evenly populated, whereas only approximately 17% of compounds in the learning set are considered soluble at the 100 mM threshold level. In consequence, the information content of the soluble set is more and more reduced, and the descriptor value range distributions of the soluble and insoluble subsets become more distinct, as confirmed by histogram analysis. According to DSE theory, this corresponds to increasing complementarity of data distributions and, therefore, results in larger DSE values. Thus, changes in the composition of the learning set enhance DSE characteristics detected at low solubility threshold levels. Since these effects are not descriptor-specific, the relative ranking of preferred descriptors is little affected, if at all.

Binary QSAR Predictions. One of our key questions has been whether descriptors selected on the basis of DSE characteristics can in fact be used to correctly predict physical characteristics of test compounds. Therefore, we have used DSE-preferred descriptors in a systematic manner to build binary QSAR models for solubility predictions. The choice of five solubility threshold values and six descriptor ranges (with the number of descriptors increasing in increments of five) resulted in the construction of 30 binary QSAR models, all of which were used for prediction of our test molecules. The obtained results are reported in Table 4, and Figure 2 monitors overall prediction accuracy as a function of the number of descriptors used to build the models. Regardless of the applied solubility threshold value, overall high prediction accuracy was achieved in our calculations, both for prediction of compounds classified as soluble and insoluble (Table 4). The best prediction accuracy was 88% when averaged over all five solubility ranges. As to be expected, based on the skewed composition of the learning set at high solubility threshold values (reducing the information content of the soluble subset, as discussed above), least accurate predictions, with an average of 70%, were obtained for the small test sets of soluble compounds at 50 mM and 100 mM (containing only 16 and 13 compounds, respectively). By contrast, at threshold values of 1 mM and 5 mM, where the compositions of both learning and test sets are numerically balanced, an average prediction accuracy of 82% was obtained for all descriptor combinations. Therefore, we can conclude that descriptors with high predictive value could be readily identified and selected on the basis of our SE-DSE analysis.

Identification of Best-Performing Descriptors. Figure 2 shows that for four of five solubility threshold values, except 100 mM, best predictions (with 88% average accuracy) were achieved using only the top five descriptors with largest DSE values (reported in Table 2). These four descriptor sets contain a total of only seven descriptors and are thus largely overlapping. They represent logP, hydrophobic van der Waals surface, hydrophobic atoms counts, and three Labute descriptors that approximate hydrophobic, polar, or charged surface areas. By contrast, descriptors accounting for hydrogen bond potential or partial charges were not required for best performance. Since we intentionally did not eliminate similar or correlated descriptors from

Table 4. Summary of Binary QSAR Predictions[a]

	A. 1 mM			B. 5 mM			C. 10 mM			D. 50 mM			E. 100 mM		
no. of descriptors	**PA**	PA_S	PA_{INS}	**PA**	PA_S	PA_{INS}	**PA**	PA_S	PA_{INS}	**PA**	PA_S	PA_{INS}	**PA**	PA_S	PA_{INS}
5	**85**	83	87	**88**	84	91	**84**	74	89	**93**	81	95	**89**	61	93
10	**81**	78	85	**82**	75	87.5	**81**	76	83	**87**	69	90	**91**	77	93
15	**79**	76	83	**77**	68	84	**84**	82	85	**85**	63	89	**90**	69	93
20	**82**	83	80	**84**	75	91	**79**	65	86	**85**	63	89	**92**	69	95
25	**80**	78	83	**80**	77	82	**83**	76	86	**84**	75	86	**91**	77	93
30	**83**	85	80	**83**	75	89	**83**	71	89	**86**	69	89	**91**	69	94

[a] Results are reported for solubility predictions on the test set, consisting of a total of 100 compounds. "PA" abbreviates prediction accuracy (as defined in the Methods section). All values are percentages. Overall prediction accuracy is shown in bold face.

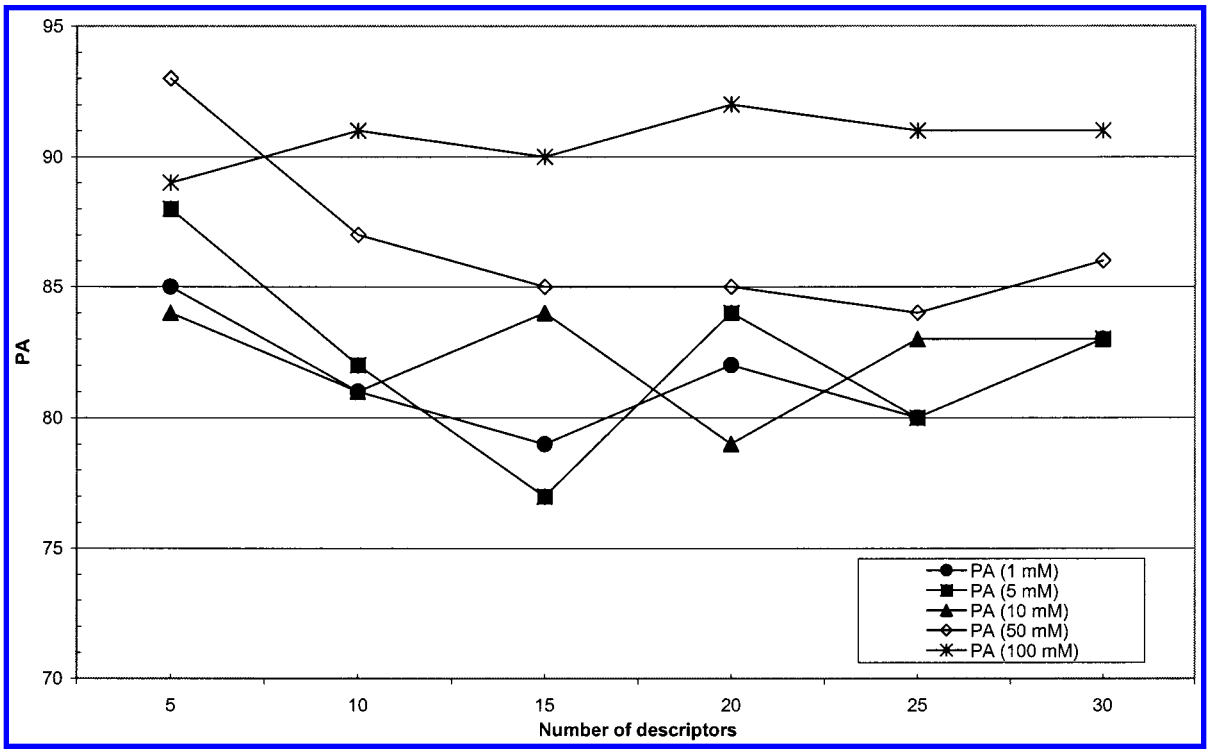

Figure 2. Accuracy of binary QSAR predictions. For each solubility threshold value, the figure reports overall prediction accuracy (PA) as a function of the number of descriptors selected to construct the binary QSAR models.

our initial pool, to have "internal controls" for entropy evaluation, it was expected to find very similar (e.g., "slogP", "logP(o/w)") or strongly correlated (e.g., "vsa_hyd", "a_hyd") descriptors in close proximity in our DSE-ranked lists. Thus, taking this into consideration, few descriptors accounting for hydrophobic character and molecular surface properties were sufficient to accurately classify compounds according to aqueous solubility over a wide solubility range, between 1 mM and 100 mM.

Figure 2 also shows that there is relatively little variation of overall prediction accuracy with increasing numbers of descriptors. However, in a number of cases, prediction accuracy decreases when more descriptors are included and drops below the 80% level (for example, for models with 10 or 15 descriptors at the 5 mM threshold). These "counter-intuitive" effects can be explained by the fact that the binary QSAR approach utilizes principal component analysis to derive its probability density function. For each descriptor combination, PCA generates a distinct reference frame for data representation that captures the entire variance created by the value ranges of the selected descriptors. PCA decorrelates the descriptor contributions, and inclusion of descriptors that substantially add to the variance may well increase the "noise" within the reference data. These "noisy" descriptors may not show strong correlation with the ones most important for accurate predictions (i.e., the top five) and, therefore, reduce the success rate by emphasizing less relevant properties in test molecules. However, overall these effects have little influence, since prediction accuracy shows less than 10% variation when the number of descriptors is incrementally increased. Figure 3 illustrates the degree of separation between soluble and insoluble compounds that can be obtained by plotting the first three principal components calculated from a top five descriptor combination. Analysis of compound positions in this PCA space confirmed that molecules with a -logS value within $\pm$ 0.5 to the solubility threshold value (10 mM for the example shown in Figure 3) map closely to the "border" between the soluble and insoluble test compound subsets.

Descriptor Selection and Prediction Accuracy in the Context of Other Solubility Calculations. Binary QSAR is best understood as a classification approach and can thus not be directly compared with conventional multiple linear regression or neural network methods. In state-of-the-art investigations on aqueous solubility using the latter methods,[12−18] r^2 correlation coefficients of 0.6−0.9 between

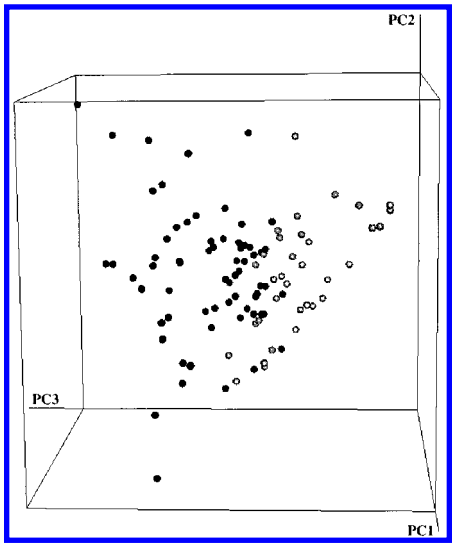

Figure 3. Visual representation of "soluble" and "insoluble" compounds in principal component space. Shown is a "descriptor space" representation of our test set compounds for the aqueous solubility threshold of 10 mM. PC −1, −2, and −3 are the first three principal components of the top five DSE-selected descriptors (see Table 3) and serve as coordinate axes. Sixty-six molecules classified as insoluble are colored black and 34 soluble molecules gray. The figure (generated with MOE[30]) illustrates that soluble and insoluble compounds are well-separated in principal component space derived from the preferred descriptor combination.

experiment and prediction have been obtained and observed errors are often within an order of magnitude. Since we have classified compounds over a relatively wide solubility range, consistently achieved prediction accuracy of >80% should in principle compare favorably with these related yet distinct investigations.

A qualitatively important aspect of our study is that very few property descriptors are sufficient to predict aqueous solubility with high accuracy. These findings are well in accord with those of Jorgensen and Duffy[12] who initially identified a set of 11 descriptors in Monte Carlo simulations and averaged these descriptors for solubility predictions. Ultimately, only five terms were required to establish a QSPR equation of high predictive value. Important descriptors included molecular volume, hydrophilic, hydrophobic, and aromatic terms, hydrogen bond descriptors, and electrostatic and van der Waals interaction energies.[12] In general, it is easy to understand which effects must in principle play a major role in determining aqueous solubility of test compounds. However, as also illustrated by our study, it is difficult to predict which of the many available descriptors, or their combinations, would best account for these effects. For example, we found a new class of surface property descriptors to be important in our analysis, whereas hydrogen bond descriptors, which have been used in other studies, were not identified as being important for accurate predictions. Taken together, the results emphasize the importance of methods to select descriptors suitable for specific applications beyond chemical intuition.

CONCLUSIONS

A major objective of our study has been to evaluate the potential of entropy-based descriptor selection. As a test case, binary QSAR-based prediction of aqueous solubility was chosen. Based on the analysis of learning sets, we have identified property descriptors with large DSE values and have shown that combinations of these descriptors accurately classify compounds according to aqueous solubility at different solubility thresholds. Since descriptors were exclusively selected on the basis of entropy values, not taking any other information into account, the results suggest that data set-specific differences in descriptor entropy correlate with detectable differences in physicochemical properties of test compounds, thus providing support for entropy-based descriptor selection methodology. We found that, at least in this case, very few descriptors were sufficient to build binary QSAR models of high prediction accuracy and that such descriptors could be confidently selected based on their DSE values. The results have some practical implications. For example, since our binary QSAR models perform very well at varying solubility threshold levels, they can readily be used to generate solubility profiles of large compound libraries. Thus, those compounds can be identified that have a high probability of solubility problems at concentrations of 5−10 mM, and this information should be very helpful in the design and interpretation of screening experiments.

ACKNOWLEDGMENT

The authors thank Douglas Kitchen and James Schermerhorn for help in the assembly of the compound data sets.

REFERENCES AND NOTES

(1) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(2) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Combin. Chem. High Throughput Screen.* **2000**, *3*, 363−372.
(3) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233−245.
(4) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, U.S., 1963.
(5) Godden, J. W.; Stahura, F. L.; Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 796−800.
(6) Godden, J. W.; Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.
(7) Godden, J. W.; Bajorath, J. Chemical descriptors with distinct levels of information content and varying sensitivity to differences between selected compound databases identified by SE-DSE analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.
(8) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between natural products and synthetic molecules by Shannon descriptor entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.
(9) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem., Intl. Ed. Engl.* **1999**, *38*, 643−647.
(10) Lipinski, C. A. Avoiding investments in doomed drugs. *Current Drug Discovery* **2001**, *1*(*2*), 17−19.
(11) Taskinen, J. Prediction of aqueous solubility in drug design. *Curr. Opin. Drug Discov. Dev.* **2000**, *3*, 102−107.
(12) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from Monte Carlo simulations. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1155−1158.
(13) Sutter, J. M.; Jurs, P. C. Prediction of aqueous solubility for a diverse set of heteroatom-containing organic compounds using a quantitative structure−property relationship. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 100−107.
(14) Mitchell, B. E.; Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 489−496.

(15) McElroy, N. R.; Jurs, P. C. Prediction of aqueous solubility of heteroatom-containing organic compounds from molecular structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.
(16) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(17) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of aqueous solubility of chemical compounds using E-state indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(18) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(19) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(20) Klopman, G.; Zhao, H. Estimation of aqueous solubility of organic molecules by the group contribution approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.
(21) Hall, L. H.; Kier, L. B. The E-state as the basis for molecular structure space definition and structure similarity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 784−791.
(22) Labute, P. Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, *7*, 444−455.
(23) Physical/Chemical Property database (PHYSPROP); Syracuse Research Corporation, SRC Environment Science Center: Syracuse, NY, 1994.
(24) Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464−477.
(25) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(26) Gao, H.; Bajorath, J. Comparison of binary and 2D QSAR analyses using inhibitors of human carbonic anhydrase II as a test case. *Mol. Divers.* **1999**, *4*, 115−30.
(27) Gao. H. Application of BCUT metrics and genetic algorithm in binary QSAR analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 402−407.
(28) Feller, W. *An Introduction to Probability Theory and its Applications*; Wiley & Sons Inc.: New York, 1950; Vol. 1.
(29) Glen, W. G.; Dunn, W. J.; Scott, D. R. Principal component analysis and partial least squares regression. *Tetrahedron Comput. Methodol.* **1989**, *2*, 349−376.
(30) MOE (Molecular Operating Environment), version 2001.01; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3.
(31) Williams, P. W. Criteria for choosing subsets to obtain maximum relative entropy. *Computer J.* **1978**, *21*, 57−62.
(32) Cringean, J. K.; Pepperrell, C. A.; Poirrette, A. R.; Willett, P. Selection of screens for three-dimensional substructure searching. *Tetrahedron Comput. Methodol.* **1990**, *3*, 37−46.
(33) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric pattern matching in files of three-dimensional structures: characterization and use of generalized valence angle screens. *J. Mol. Graph.* **1991**, *9*, 203-217.
(34) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, NY, U.S., 1997.

CI010243Q