An Information-Theoretic Approach to Descriptor Selection for Database Profiling and QSAR Modeling

Jeffrey W. Godden[a] **and Jürgen Bajorath**[a, b]*

[a] Albany Molecular Research, Inc., Bothell Research Center (AMRI-BRC), 18804 North Creek Pkwy, Bothell, Washington 98011, and
[b] Department of Biological Structure, University of Washington, Seattle, WA 98195

Abstract

In order to rationalize the selection of molecular descriptors for QSAR and other applications, we have adapted the Shannon entropy concept that was originally developed in digital communication theory. The approach has been extended to facilitate the large-scale analysis of molecular descriptors and their information content in diverse compound databases. This has enabled us to identify descriptors with consistently high information content. Furthermore, it has been possible to select descriptors that are sensitive to systematic property differences in diverse compound collections (synthetic compounds, natural products, drug-like molecules, or drugs) and, in addition, to quantify such database-specific differences. Selection of descriptors based on information content has been proven useful for binary QSAR analysis. In this review, we describe the principles of entropy-based descriptor selection and discuss different applications.

Introduction

Descriptors of molecular structure and properties are important components of many investigations in computational medicinal chemistry and chemoinformatics [1]. For example, descriptors are essential for QSAR-type applications, definition of chemical spaces for diversity or similarity analysis, or virtual screening. Given the fact that literally hundreds of in part very different descriptors are available [1, 2], it is important to better understand the sensitivity of diverse molecular descriptors to chemical information and to rationalize descriptor selection. In this context, a number of questions are relevant. For example, which property descriptors capture significant amounts of chemical information in diverse compound databases? Can we compare descriptor distributions in detail and quantify their differences? Which descriptors are sensitive to systematic chemical differences between molecular data sets? These and other questions prompted us to explore novel ways of systematically analyzing the database variability of molecular descriptors. In order to do so, it became necessary to capture chemical information indirectly, which means irrespective of differences in the properties represented by various descriptors, their units, or value ranges. Considering this task, variability or frequency analysis could be related to entropy calculations, at least in principle, which suggested to us the design and implementation of a suitable entropy-based approach. About a decade ago, relative entropy calculations were first introduced in chemical database analysis to study angle and distance ranges of compounds or to establish the frequency of occurrence of substructures [3, 4]. However, we focused our attention on some information-theoretic approaches, since they were thought to be most suitable for systematic and large-scale comparison of descriptor variability.

* To receive all correspondence at the AMRI-BRC address, Tel.: (425) 424-7297, fax: (425) 424-7299, e-mail: jurgen.bajorath@albmolecular.com.

Key words: Molecular descriptors, information content, Shannon entropy, database profiling, binary QSAR

Abbreviations: QSAR, Quantitative Structure-Activity Relationship; SE, Shannon Entropy; DSE, Differential Shannon Entropy; sSE, Scaled Shannon Entropy; sDSE, Scaled Differential Shannon Entropy; KL, Kullback-Leibler function; bQSAR, Binary QSAR; 2D, Two-Dimensional; 3D, Three-Dimensional; MOE, Molecular Operating Environment; ACD, Available Chemicals Directory; C&H, Chapman and Hall Dictionary of Natural Products; CMC, Comprehensive Medicinal Chemistry Database; SYNTH, Synthline Database; PHYSPROP, Physical/Chemical Property Database; vdW, van der Waals.

Methods

Basic concepts. Our initial goal was to quantitatively describe the distribution of molecular descriptors in compound databases. Therefore, two similar yet distinct concepts from information theory were considered. One was

Shannon entropy (SE), originally introduced by Claude Shannon in the early 1960s for application in digital communication theory [5] and the other the Kullback-Leibler information number, a function used to quantify the similarity between a true data distribution and a statistical model [6]. With modifications, both approaches could be adapted to reduce chemical data distributions to their information content.

Shannon entropy is defined as

$$SE = -\Sigma p_i \log_2 p_i$$

with 'p' being the probability of a data point or count 'c' to fall within a specific data range 'i'. Therefore, 'p' is obtained as

$$p_i = c_i/\Sigma c_i$$

In this formulation, the logarithm to the base two can be rationalized a binary detector of counts and a scale factor of information content. Essentially, it captures the information whether or not a count appears in a specific data interval.

The Kullback-Leibler function, on the other hand, is defined as:

$$\mathrm{KL} = \sum \mathrm{g_x(x)} \log\left[\frac{\mathrm{g_x(x)}}{\mathrm{h_x(x)}}\right]$$

where 'g_x' is a true data distribution and 'h_x' a statistical model to approximate g_x. Thus, in essence, KL calculates a relative entropy term. Considering the characteristics of SE and KL, we concluded that SE would be better suited for comparing descriptor distributions, for two reasons. Firstly, the value of KL depends on which data distribution is considered the observed or true distribution and, secondly, KL is not defined if the model distribution has zero probability in a given data interval (which is often the case for descriptors having discrete value ranges). Consequently, we focused on adapting the SE metric.

Shannon entropy approach. A major feature of SE analysis is that diverse descriptor distributions can be reduced to information content as long as the data presentation is uniform. This can be conveniently achieved by representing distributions of descriptor values in histograms with fixed bin number. Thus, in our implementation of the SE approach, we initially divided any descriptor value range into the same number of data intervals [7]. In practice, these histograms are obtained by calculating the values of a given descriptor for all compounds in a database and plotting their frequency of occurrence against the observed value range. Then, the number of data counts per interval is determined and transformed into probabilities of data occurrence, which are used to calculate SE values. Figure 1 shows model histograms of data distributions having minimum, intermediate, and maximum information content, as determined by SE calculations. Information content may range from zero for a distribution with only a single value to a maximum of the logarithm to the base two of the total number of histogram bins. For example, if 100 bins are used, the maximum possible SE value is $\sim$6.6.

A bin number-independent SE value can be obtained by scaling (sSE), i.e., by dividing the observed SE by the maximum possible SE value for the number of bins used (SE divided by the logarithm to the base two of the total number of bins):

$$sSE = SE/\log_2 \text{ (bins)}$$

Accordingly, sSE values range from zero to one (maximum information content). Figure 2 shows example histograms for molecular descriptors having significantly different information content. Broad value distributions correspond to large sSE values, whereas narrow distributions yield smaller values, as one would expect.

Differential Shannon entropy. An important question has been how one can best compare descriptor database distributions in quantitative terms. As illustrated in Figure 3, a simple comparison of sSE values is not sufficient for this purpose because this does not take differences in value range occupancy into account. Therefore, we have introduced an extension of the SE concept, differential Shannon entropy (DSE) [8], which is defined as

$$DSE = SE_{AB} - (SE_A + SE_B)/2$$

where 'SE_A' is the SE value of a descriptor in database A, 'SE_B' the corresponding value in database B, and 'SE_{AB}' the SE value calculated for the combination of the two databases. In analogy to sSE, DSE values can also be scaled and made independent of histogram bin numbers. Scaled DSE is defined as

$$sDSE = DSE/\log_2 \text{ (bins)}$$

A key feature of DSE is that this formulation takes both differences in the variability and value range distributions of descriptors into account. This is important because combinations of descriptor distributions are not necessarily the sum of single distributions. Combining such distributions requires renormalization of the data. In this regard, DSE represents the difference between the renormalized histogram of both distributions and the average of the independent distributions. Figure 4 illustrates that DSE mirrors differences in information content that include complementary features of the compared distributions. If combining and renormalizing two distributions does not increase overall variability, negative DSE values can also be observed. All programs required for SE and DSE analysis were written in Perl.

Binary QSAR. One of the initial goals of descriptor information content analysis was to provide a rationale for the selection of descriptors for specific applications in binary QSAR (bQSAR), a probabilistic QSAR-like approach based on Bayes' Theorem [9] and developed by Labute

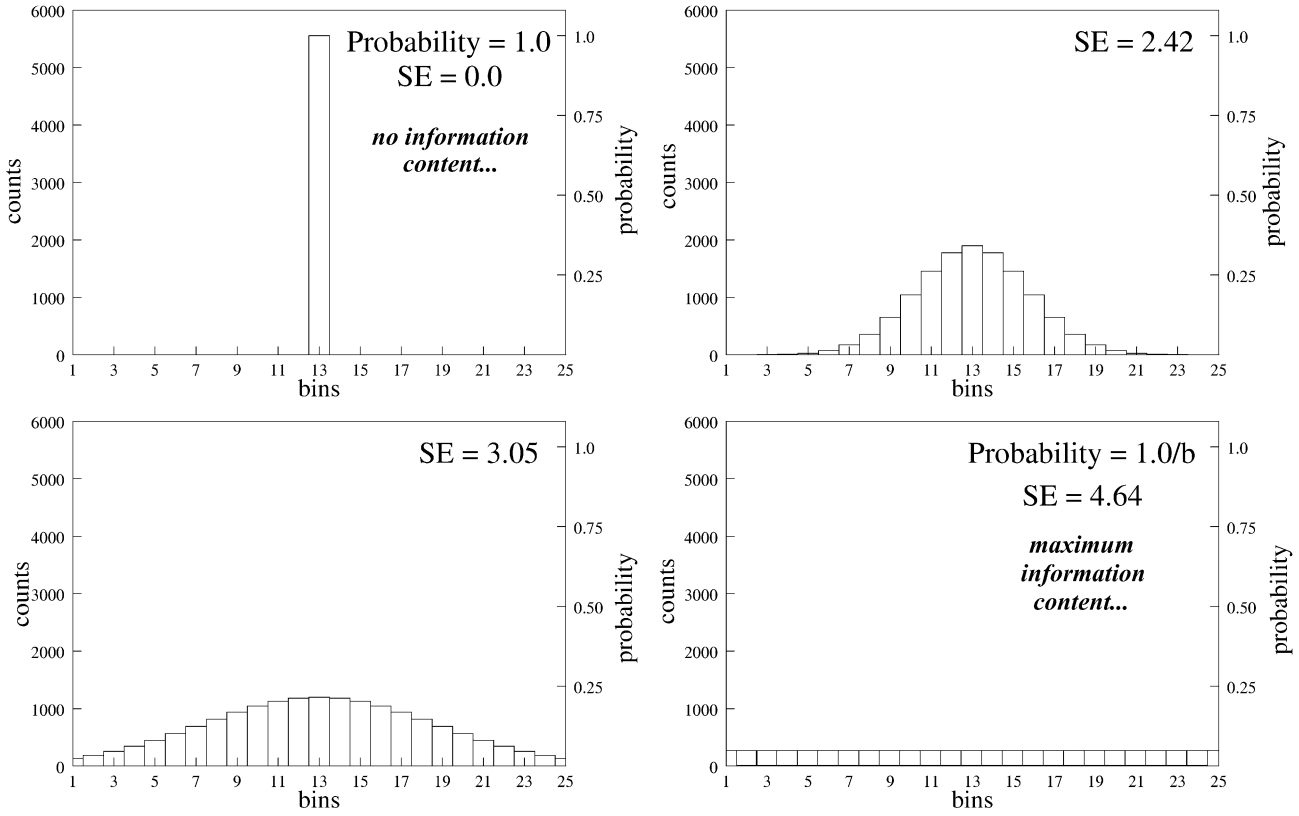

Figure 1. Model data distributions and corresponding SE values. Data representation is uniform, which means that each data range is evenly divided into 25 bins. Therefore, the SE values shown here can be directly compared. For 25 histogram bins, the maximum possible SE value is approximately 4.6, as represented by the distribution with maximum information content on the lower right. Here a probability of 1.0/b means that 1.0 is divided by the total number of histogram bins.

[10]. Starting with the analysis of learning sets, bQSAR correlates structural features or properties of molecules that are captured by descriptors with a binary (yes/no or 1/0) formulation of biological activity or other molecular features. Each investigated descriptor combination is transformed into a specific probability density function and used as a model to predict the binary state of molecules in test sets. Since the probability function produces continuous values between 0 and 1, a cut-off must be defined (typically 0.5) to discriminate between binary states. The bQSAR calculations reported herein were carried out with the Molecular Operating Environment (MOE, Chemical Computing Group Inc., 1255 University Street, Montreal, Quebec, Canada, H3B 3X3).

Descriptors and databases. For descriptor analysis, a pool of approximately 150 molecular descriptors was used, as described previously [8], including bulk property descriptors and various types of 2D or implicit 3D descriptors (e.g., surface area approximated from 2D molecular representations). All descriptor values were calculated with MOE. Molecular descriptors reported in the Results section are explained in Table 1. In contrast to numerical descriptors, two-state descriptors such as structural fragments or keys [11] that are either present or absent in a molecule cannot be subjected to SE and DSE analysis in a meaningful way. This is the case because two-state descriptors lack value range distributions and have therefore no information content that can be quantified by SE calculations.

Descriptor comparisons were primarily carried out in four compound databases, the ACD (Available Chemicals Directory, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, USA; containing ~200 000 molecules), C & H (Chapman and Hall Dictionary of Natural Products, CRC Press LLC, 2000 NW Corporate Blvd., Boca Raton, FL 33431, USA; ~120 000 entries), CMC (Comprehensive Medicinal Chemistry Database, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577, USA; ~8000 molecules), and SYNTH (Synthline Drug Database, Prous Science, Provenza 388, 08025 Barcelona, Spain; ~4000 compounds). For prediction of aqueous solubility, a subset of PHYSPROP (Physical/Chemical Property Database, Syracuse Research Corporation, 6225 Running Ridge Road, North Syracuse, NY 13212, USA) was used.

Results

SE and DSE calculations. In general, we found that descriptor entropies varied greatly in the compound data-

Table 1. Molecular descriptors reported in this study.

Abbreviation	Definition
SlogP_VSA2	approx. van der Waals (vdW) atomic surface with $-0.2 < \log P \leq 0.0$ [12]
SlogP_VSA6	approx. vdW surface with $0.2 < \log P \leq 0.25$
SlogP_VSA7	approx. vdW surface with $0.25 < \log P \leq 0.3$
SMR_VSA3	approx. vdW surface with molar refractivity is $0.35 < R_i \leq 0.39$ [12, 26]
SMR_VSA5	approx. vdW surface with molar refractivity is $0.44 < R_i \leq 0.485$
SMR_VSA6	approx. vdW surface such with molar refractivity is $0.485 < R_i \leq 0.56$
a_ICM	entropy of element distribution in the molecule
SMR	molar refractivity
mr	alternative formulation of molar refractivity
logP(o/w)	octanol/water partition coefficient
SlogP	alternative formulation of the octanol/water partition coefficient
a_nC	number of carbon atoms in a molecule
a_hyd	number of hydrophobic atoms
b_rotR	fraction of rotatable bonds
b_1rotR	fraction of rotatable single bonds
b_heavy	number of bonds between heavy atoms.
b_single	number of single bonds
balabanJ	Balaban's topological connectivity index [27]
chi1v	atomic valence connectivity index (order 1)
chi0_C	sum of the inverse square root of heavy atoms bonded to each atom
chi0v_C	sum of the inverse square root of a valence electron function.
chi1_C	sum of the inverse sq. r. of cross terms of valence electron function
chi1v_C	carbon valence connectivity index (order 1)
weinerPol	half the sum of all the distance matrix entries with a value of 3 [28]
vsa_hyd	approx. van der Waals surface area of hydrophobic atoms.
PEOE_VSA -1	vdW surface area with atomic partial charge $-0.10 \leq q < -0.05$ [12, 29]
PEOE_VSA $+0$	vdW surface area with atomic partial charge $0.00 \leq q < 0.05$
PEOE_VSA $+1$	vdW surface area withtomic partial charge $0.05 \leq q < 0.10$
PEOE_VSA_HYD	polar vdW surface area of hydrophobic atoms
PEOE_VSA_FHYD	fractional polar vdW surface area of hydrophobic atoms
PEOE_VSA_FPOL	fract. polar vdW surface area with abs. value of partial charge >0.2
PEOE_VSA_NEG	total negative vdW surface area
PEOE_VSA_FPPOS	fract. positive vdW surface with partial charge >0.2/total surface area
PEOE_VSA_FPNEG	fract. negative vdW surface partial charge <-0.2/the total surface area
PEOE_RPC+	largest positive atomic partial charge divided by the positive sum
VadjEq	function of a logarithm to the base of two of adjacency map
VadjMa	one plus the log two of the number of heavy-heavy bonds
VdistEq	sum of log two of distance matrix entries minus a function of distance matrix entries of a common value
VdistMa	sum of log two of distance matrix entries minus funct. of shortest path
zagreb	sum of squares of the number of heavy atoms bonded to each atom

bases we analyzed. We consistently identified descriptors with high, medium, and low information content, which often displayed significant database-dependence [7, 8]. Some descriptors with relatively complex design such as deduced molecular surface area descriptors with mapped physical properties [12] had high information content in all compound databases. On the other hand, DSE calculations revealed that rather simple descriptors such as counts of specific atom types or bonds (e.g., the number of aromatic atoms or hydrogen bond acceptors in a molecule) showed significant differences in pair-wise database comparisons (or, in other words, highly complementary value distributions) [8]. In addition, we found that different types of molecular descriptors were most responsive to intrinsic chemical differences between databases, for example, when comparing ACD and C&H (i.e., synthetic compounds versus natural products) [8, 13] or ACD and CMC (i.e., synthetic compounds versus drug-like molecules) [8]. Based on these observations, we attempted to derive a more generally applicable scheme for the classification of molecular descriptors according to information content and database-dependent differences. Ultimately, this led to the concept of SE-DSE analysis, as described in the following.

s**SE-sDSE classification.** One of the key questions for the design of this descriptor classification scheme was what level of information content could generally be considered as high. In order to address this question, we surveyed sSE and sDSE values for 143 descriptors in the ACD, C&H, CMC, and SYNTH databases [14]. In these calculations, a total of 495 non-zero sSE values were obtained and graphically analyzed in combination, as shown in Figure 5. The observed sSE distribution was bimodal with a Gaussian-like tendency towards high values. Considering the overall shape of this distribution, we defined an sSE value of 0.3 as a

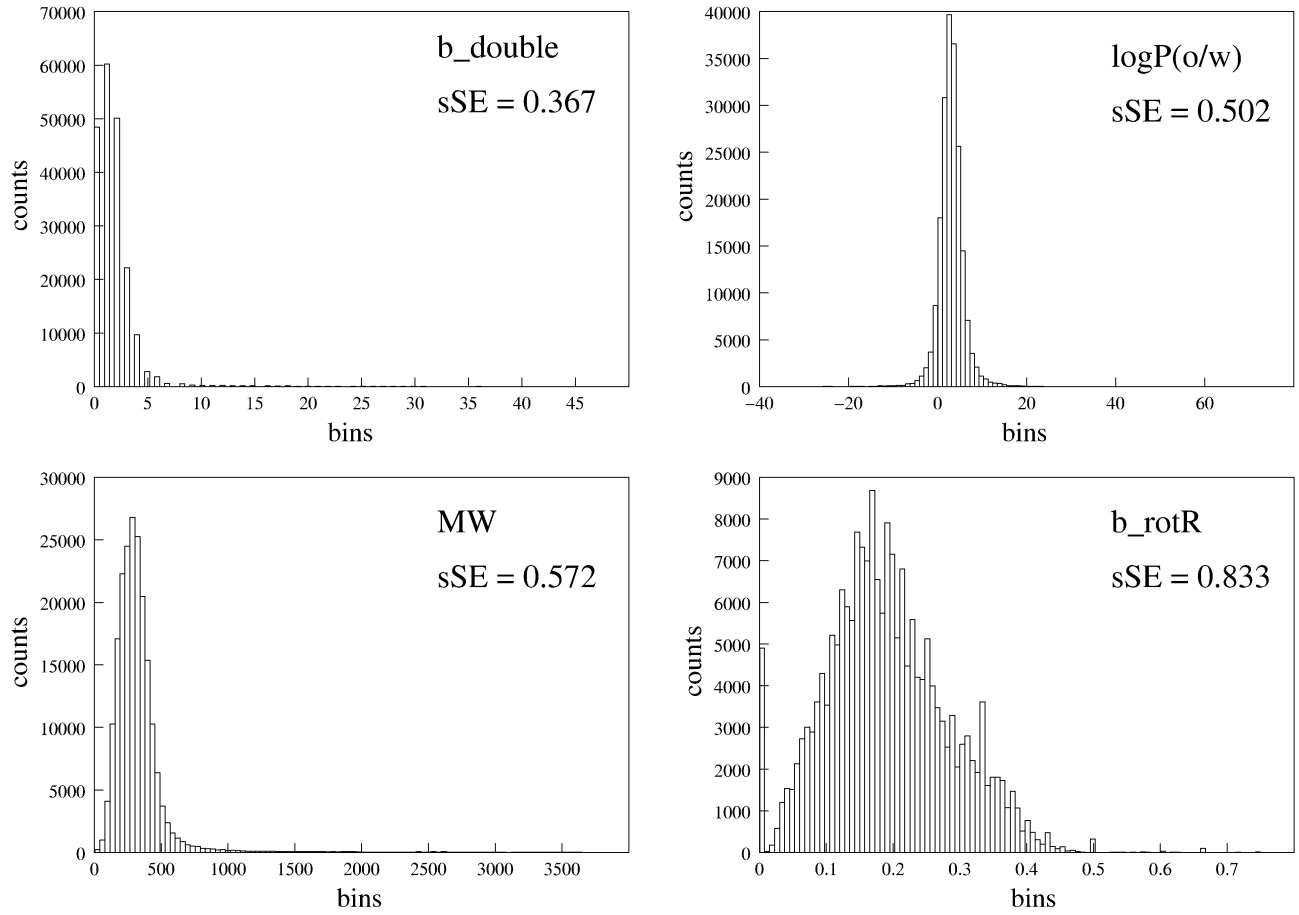

Figure 2. Representative histograms of descriptor distributions in the Available Chemicals Directory. Four descriptor distributions with increasing (bin-independent) sSE values are shown. The descriptor b_double counts the number of double bonds in a molecule, logP(o/w) is the octanol-water partitioning coefficient, MW stands for molecular weight, and b_rotR accounts for the fraction of rotatable bonds in a molecule.

general threshold for low sSE and an sSE value of 0.6 as a threshold for high sSE. Values between 0.3 and 0.6 were considered intermediate. In addition, sDSE calculations were carried out for all descriptors and pair-wise comparisons of the four databases, producing a total of 858 values. The combined sDSE distribution is also shown in Figure 5. Since sDSE calculation is an averaging operation, this distribution resulted from deviations of mean values and was therefore also Gaussian-like. When a normal curve was fitted to the sDSE distribution, a standard deviation (or sigma value) of 0.026 sDSE units was obtained. We considered sDSE values outside of this one sigma limit as high sDSE and values inside one sigma as low sDSE.

The determination of these sSE and sDSE threshold values made it possible to define four basic sSE-sDSE categories (high-high, high-low, low-high, and low-low) for the comparison of descriptor database variability [14]. Of these, the high-high and high-low categories are the most interesting ones because they contain descriptors having high information content. In the six pair-wise database comparisons we carried out, only 11 of 143 descriptors were found to belong to the high-high sSE-sDSE category, as reported in Table 2. None of these descriptors commonly occurred in all database comparisons. Some of the calculations gave rather unexpected results. For example, in the high-high category was a very simple descriptor counting the number of single bonds in a molecule, and this descriptor detected unexpected differences between drug-like molecules and known drugs. It also responded to intrinsic differences between synthetic and naturally occurring molecules (which might be more intuitive). In general, descriptors belonging to the high-high category have consistently high information content in the compared databases but significantly different value distributions. Therefore, these information-rich descriptors are most sensitive or responsive to intrinsic differences between the synthetic, natural, or drug-like molecules we compared and best reflect their diversity. In addition, we identified 22 descriptors belonging to the high-low category that had consistently high sSE values in the four databases and low sDSE values in each of the six database comparisons. These descriptors are listed in Table 3. These findings confirmed that information-rich descriptors are not necessarily sensitive to compound class-specific features or systematic chemical differences. In

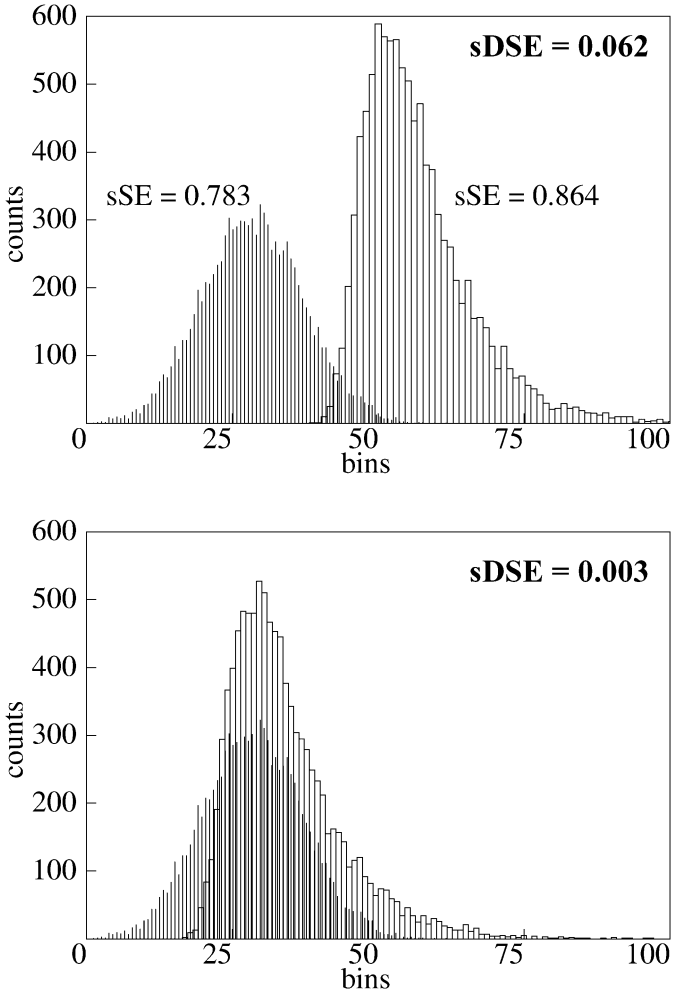

Figure 3. sDSE as a function of value range occupancy. Two hypothetical descriptor distributions with significant information content are shown. In both cases, the information content of each distribution is the same but their relative value range occupancy differs. Simple subtraction of their sSE values would produce a difference of 0.081 in both situations (top and bottom). However, in the top diagram, the value ranges of the two distributions differ significantly, yielding a large sDSE value. By contrast, in the bottom diagram, the value ranges closely overlap and, consequently, the resulting sDSE value is small.

fact, in our analysis, descriptors belonging to the high-low sSE-sDSE occurred much more frequently than high-high descriptors. The so derived information content-dependent classification scheme provided a basis for rational selection of descriptors for QSAR applications.

Distinguishing between compounds from different sources. In a first application of descriptor entropy analysis, we attempted to systematically distinguish between synthetic molecules and natural products. This analysis was suggested by our observations that the variability of specific descriptors varied significantly in ACD and C&H, as revealed by SE calculations [13]. Furthermore, synthetic and naturally occurring molecules were considered an interesting test case, since relatively little was known about intrinsic and quantitative differences between these classes of molecules. Only very few studies had addressed these issues. For example, a statistical analysis was available that compared the distributions of molecular fragments, functional groups, and properties in natural, synthetic, and drug-like molecules and revealed some systematic chemical differences [15]. These included, for example, the on average higher molecular weight of natural products and their increased oxygen content relative to synthetic molecules that are generally richer in nitrogen-containing functional groups [15]. Another subsequent analysis reported overall similar findings and revealed that core structures of natural products and drug molecules have little overlap, whereas their pharmacophore patterns display distinct similarity [16].

For our analysis, we selected from SE calculations a number of property descriptors that were variably set in ACD and C&H plus variable structural keys, as identified by their relative frequency of occurrence in these databases. Using alternative descriptor combinations, several bQSAR models were derived to systematically distinguish between randomly assembled test sets of synthetic and natural molecules (with or without specific biological activity) [13]. In these calculations, different bQSAR models and descriptor sets achieved greater than 80% prediction accuracy. These descriptor combinations consisted of, on average, less than ten molecular descriptors or structural keys. The best-performing bQSAR model was derived from only seven descriptors including three structural keys (accounting for hydroxyl groups, oxygen atoms attached to a ring, and double bonds, respectively) and four rather simple 2D descriptors (the number of hydrogen atoms, single or aromatic bonds, and the element distribution in a molecule). When applied to different test sets, some of which consisted of synthetic and natural compounds with specific activity, the model consistently produced between 81% and 93% prediction accuracy [13].

Prediction of aqueous solubility. Since the bQSAR classification of natural and synthetic molecules was based only on SE (but not DSE) calculations and took structural keys into account (that are not amenable to SE-DSE analysis), we subsequently attempted to predict aqueous solubility of organic molecules as another test case. Here DSE-based selection of only numerical descriptors was applied for bQSAR modeling. For this analysis, an important question was whether or not significant differences in descriptor information content could be correlated with measurable changes in molecular properties such aqueous solubility.

Limited or excessive solubility of database compounds has been recognized as a major error source in biological screening and has thus become a topic of intense computational research, aiming at the derivation of reliable models for solubility predictions [17]. However, our bQSAR analysis conceptually differed from other predictive approaches that are prevalent in the field, in particular, QSPR models [18], neural networks simulations [19], and additive

Table 2. Descriptors of the high-high sSE-sDSE category.

Databases 1/2	Descriptor	|sDSE|	sSE 1	sSE 2
ACD/C&H	a_ICM	0.042	0.808	0.704
ACD/CMC	PEOE_VSA+0	0.027	0.634	0.697
ACD/SYNTH	PEOE_RPC+	0.050	0.739	0.638
	PEOE_VSA+0	0.032	0.634	0.701
	PEOE_VSA_FPOL	0.032	0.853	0.811
	a_ICM	0.033	0.808	0.741
	balabanJ	0.033	0.694	0.626
C&H/CMC	SMR_VSA3	0.062	0.670	0.534
	SMR_VSA6	0.050	0.582	0.703
	SlogP_VSA2	0.030	0.644	0.576
	weinerPol	0.035	0.727	0.655
C&H/SYNTH	SMR_VSA6	0.056	0.617	0.742
	SlogP_VSA7	0.059	0.539	0.670
	b_single	0.029	0.689	0.629
	balabanJ	0.029	0.735	0.676
	weinerPol	0.036	0.727	0.653
CMC/SYNTH	PEOE_VSA_FPOL	0.113	0.829	0.811
	b_single	0.029	0.652	0.549

Reported are absolute sDSE values. Data were taken from reference 14.

Table 3. Descriptors with consistently high sSE values and low sDSE values in all database comparisons.

Descriptor	low |sDSE| (<0.026)
PEOE_VSA_FPNEG	0.003
PEOE_VSA_NEG	0.006
PEOE_VSA_FHYD	0.012
SMR_VSA5	0.012
a_hyd	0.012
chi1_C	0.012
b_rotR	0.014
a_nC	0.015
b_1rotR	0.015
PEOE_VSA_HYD	0.017
VadjEq	0.017
chi0_C	0.017
chi0v_C	0.018
PEOE_VSA_FPPOS	0.020
VdistEq	0.020
VdistMa	0.020
b_heavy	0.020
chi1v_C	0.020
vsa_hyd	0.020
VAdjMa	0.021
PEOE_VSA+1	0.023
zagreb	0.024

Reported are absolute sDSE values. Data were taken from reference 14.

group contribution methods [20]. Rather than calculating explicit solubility values, our study was designed to predict whether test sets of compounds would be soluble or not at given solubility threshold values [21]. This design reflected the major opportunities and limitations of the bQSAR approach. Compared to conventional QSAR methods, an advantage of bQSAR is that it can be applied to efficiently process very large and diverse datasets, even if data values in learning sets are only approximate. However, as a probabilistic method, the major drawback of bQSAR is that it is not possible to compute exact values for molecular features such as solubility or activity. Therefore, our analysis was limited to the prediction of solubility relative to given threshold values.

For solubility calculations, we randomly divided 650 molecules from the PHYSPROP database with known aqueous solubility into a global training set of 550 and a test set of 100 molecules. Then both training and test molecules were divided into soluble and insoluble subsets according to five solubility threshold values (1, 5, 10, 50, and 100 mM), thereby producing a total of five different training and test sets. Values of 148 descriptors were calculated for the 650 molecules and our SE-DSE metric was applied to determine which descriptors had largest DSE values for the comparison of the soluble and insoluble subsets of each training set. For each set, an increasing number of descriptors with largest sDSE values was used to build different bQSAR models that were then applied to predict the solubility of the 100 test molecules relative to the five threshold values. As representative examples, Table 4 reports the top ten descriptors with largest sDSE values for two solubility threshold values and training sets. Table 5 summarizes the results of all bQSAR predictions. The most variable descriptors differed in each of the top ten lists (for the solubility five threshold values). However, some descriptors that were chemically intuitive to account for solubility differences such as the water/octanol partition coefficient (logP(o/w)) or the number of hydrophobic atoms in a molecule (a_hyd) were consistently found among the descriptors with largest sDSE values (Table 4). Importantly, five or ten descriptors with largest sDSE values for

Table 4. Descriptors with largest sDSE values at different solubility threshold values.

5 mM				10 mM			
Descriptor	|sDSE|	sSE(s)	sSE(ins)	Descriptor	|sDSE|	sSE(s)	sSE(ins)
a_hyd	0.116	0.603	0.841	PEOE_VSA_NEG	0.125	0.496	0.754
logP(o/w)	0.114	0.647	0.754	a_hyd	0.125	0.582	0.841
SlogP	0.111	0.582	0.733	logP(o/w)	0.123	0.603	0.754
vsa_hyd	0.103	0.647	0.776	SlogP	0.119	0.560	0.733
PEOE_VSA_NEG	0.099	0.560	0.754	PEOE_VSA – 1	0.114	0.517	0.819
PEOE_VSA_HYD	0.097	0.647	0.776	chi1v	0.111	0.582	0.841
chi1v	0.093	0.603	0.841	vsa_hyd	0.108	0.625	0.776
chi1v_C	0.091	0.582	0.776	PEOE_VSA_HYD	0.103	0.603	0.776
SMR	0.091	0.647	0.862	chi1v_C	0.103	0.560	0.776
chi1_C	0.089	0.647	0.841	SMR	0.103	0.625	0.862

For solubility threshold values of 5 and 10 mM, the ten descriptors with largest absolute sDSE values are shown (obtained by sSE comparison of the respective soluble and insoluble subsets of the training set). sSE(s) and sSE(ins) report the scaled SE values of descriptors in the soluble and insoluble compound subsets, respectively. Adapted from reference 21.

Table 5. Overall accuracy of solubility predictions by bQSAR.

Number of descriptors	Prediction accuracy (%)				
	1 mM	5 mM	10 mM	50 mM	100 mM
5	85	88	84	93	89
10	81	82	81	87	91
15	79	77	84	85	90
20	82	84	79	85	92
25	80	80	83	84	91
30	83	83	83	86	91

Predictions were carried out on five test sets, each consisting of the same 100 molecules (with known aqueous solubility). Dependent on the given solubility threshold level, the ratio of compounds classified as soluble or insoluble varied (i.e., increasing threshold values correspond to smaller fraction of soluble test molecules). For 1 and 5 mM, the soluble and insoluble subsets of test molecules were about equally populated. The overall prediction accuracy was calculated as an average of prediction accuracies for the soluble and insoluble subsets. Data were taken from reference 21.

comparison of molecules above or below each solubility threshold were sufficient to yield high prediction accuracy between 80% and 90% (over the entire solubility range) (Table 5). In these calculations, the top five descriptors with largest sDSE alone achieved an average prediction accuracy of 88% [21]. These findings indicated that differential descriptor entropies correlated with differences in physicochemical molecular properties and supported the predictive value of information content-based descriptor selection.

Discussion

An information-theoretic methodology has been developed for the systematic analysis of descriptor distributions in various compound databases. One of major the goals was to provide a sound basis for descriptor selection beyond experience and chemical intuition. Since entropy calculations reduce database distributions of descriptors to their information content, descriptor characteristics can be quantitatively compared, even if the targeted properties, physical units, and value ranges differ. Analyzing and comparing descriptor settings in diverse compound databases by entropy calculations makes it also possible to generate database profiles and reveal systematic chemical differences. From this point of view, the approach presented herein is conceptually related to statistical analyses of property distributions in databases [22, 23]. Although methodologically distinct, these approaches have similar goals. Furthermore, entropy analysis as presented herein is certainly not the only possible approach to systematic descriptor selection. Database characteristics revealed by property distribution analysis can point at descriptors that are most relevant to capture similarities and differences. In addition, machine learning techniques can be applied to select preferred descriptors from large pools. For example, for QSAR or compound classification, genetic algorithms can be implemented to automatically select descriptor combinations that satisfy pre-defined fitness functions. However, a major attraction of entropy-based descriptor selection is that systematic differences between compound sets or databases can be quantified, even if these differences are small and otherwise difficult to detect.

Systematic SE calculations on a relative large number of numerical descriptors demonstrated that descriptor information content is in general (but certainly not always) approximately correlated with the complexity of their designs. Among the most complex and information-rich descriptors, we found were those that combined information from two or more descriptor types, for example, molecular surface area terms and mapped physico-chemical properties. Although we have thus far only analyzed descriptors that can be calculated from 2D molecular representations of molecules, SE calculations can be readily applied to determine the absolute and relative information content of 3D descriptors, as is the case with any other numerical descriptors.

Going a step further, DSE calculations often also revealed some intuitive trends when comparing different databases. For example, on average, DSE values were smaller for the

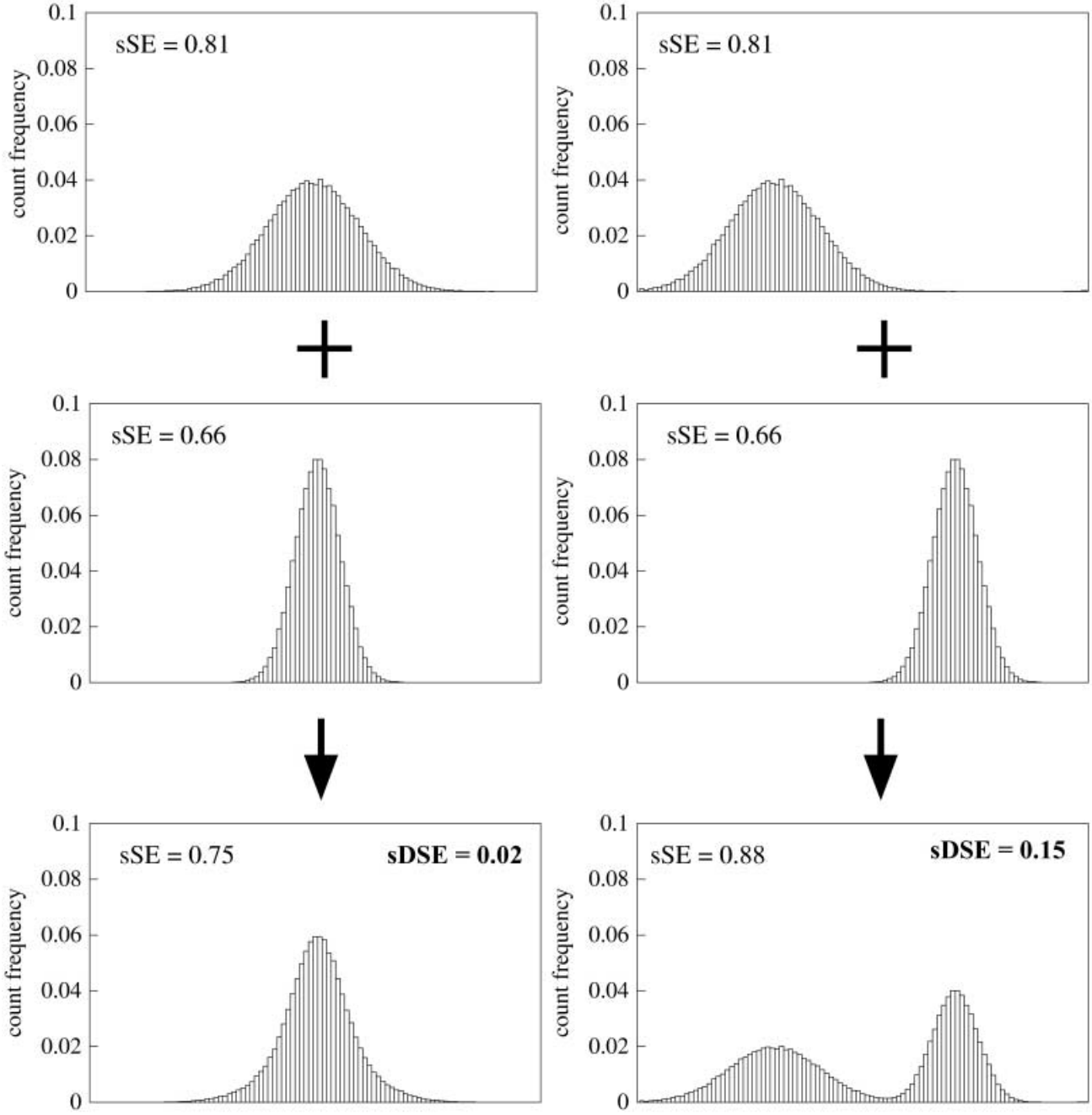

Figure 4. Complementary data distributions produce large sDSE values. Value distributions of two descriptors (left and right) in two compound databases (top and middle diagrams) are shown and the combined and renormalized distributions (bottom histograms). sDSE is calculated as the difference between the renormalized histogram of both databases binned together (bottom) and the average of their independent histograms. The more distinct or complementary the original data distributions are, the larger become the sDSE values.

ACD/CMC comparison than for comparison of ACD or CMC with natural products (C&H), which indicated that many natural products are chemically more distinct from synthetic compounds than drug-like molecules. To give another example, a simple descriptor accounting for nitrogen atoms in a molecule displayed the largest DSE value in our ACD/C&H comparison, which reflects the prevalence of amide groups in synthetic compounds as opposed to naturally occurring molecules (that are richer in oxygen). Also evident were known differences in aromatic character of synthetic and natural molecules and in halogen content, which is a characteristic feature of many ACD and CMC compounds.

Ultimately, SE-DSE analysis was established to systematically detect and quantify differences in descriptor database distributions, even if they were subtle. Considering the

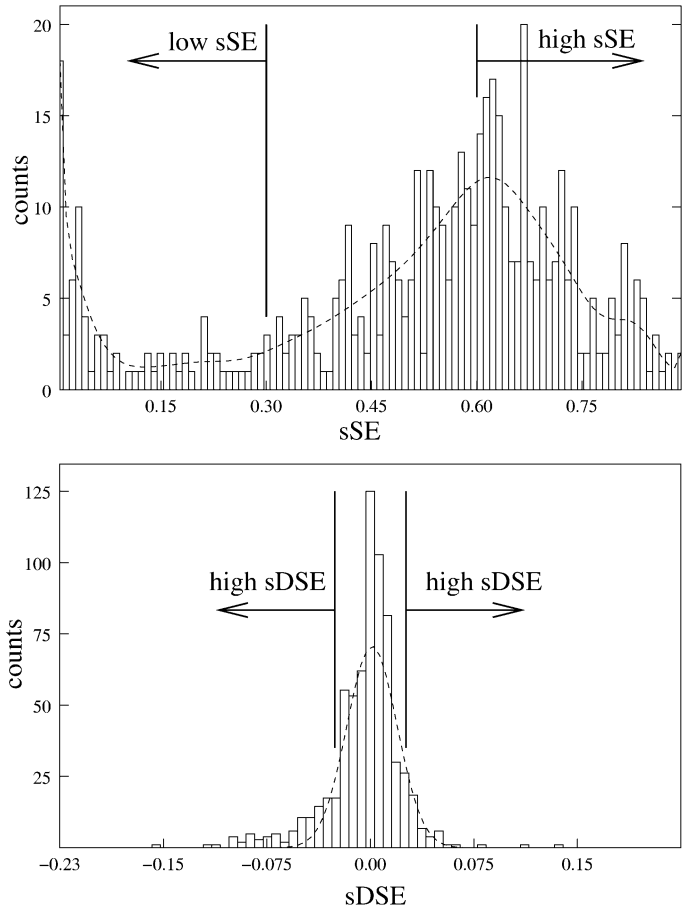

Figure 5. Determination of sSE and sDSE threshold values for descriptor classification. At the top, a histogram combining 495 non-zero sSE values from ACD, C&H, CMC, and SYNTH is shown. The bimodal nature of the distribution is captured by a dashed curve. The chosen sSE threshold values of 0.30 for low sSE and 0.60 for high sSE are indicated. The bottom graph shows the histogram representation of 858 sDSE values obtained by pairwise comparison of four database distributions of 143 descriptors. sDSE values can be positive or negative. The sDSE threshold value was defined by the one sigma limit of the fitted normal curve (dashed line). Accordingly, descriptors were designated high sDSE if their absolute sDSE value was greater than 0.026. The representation was adapted from reference 14.

results of our calculations, some guidelines for descriptor selection could be formulated. For example, descriptors belonging to the high-high sSE-sDSE category are most likely to detect systematic chemical differences between compound sets and to differentiate between compounds from diverse sources. However, many descriptors having high information content belong to the high-low sSE-sDSE class and do not detectably respond database-specific features. We conclude that these descriptors are a preferred choice for similarity searching across diverse databases. For example, this would be the case when searching known drugs or leads against synthetic databases or when trying to identify synthetic mimics of natural products with specific activity [24].

Descriptor information content analysis has found meaningful applications in QSAR-like analysis and other computational investigations, most recently in the development of a novel partitioning method [25]. In a bQSAR study of aqueous solubility we could demonstrate that differences in relative descriptor information content, as revealed by DSE calculations, correlated with differences in physico-chemical properties of test molecules, which provided substantial support for the DSE-based descriptor selection approach. Taken together, the findings discussed herein suggest that concepts from information theory should merit further investigation in computational chemistry and chemoinformatics research.

References

[1] Xue, L., and Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening, *Combin. Chem. High Throughput Screen., 3*, 363–372 (2000).

[2] Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey, *J. Chem. Inf. Comput. Sci., 40*, 195–209 (2000).

[3] Cringean, J. K., Pepperrell, C. A., Poirrette, A. R., and Willett, P. Selection of screens for three-dimensional substructure searching, *Tetrahedron Comput. Methodol., 3*, 37–46 (1990).

[4] Poirrette, A. R., Willett, P., and Allen, F. H. Pharmacophoric pattern matching in files of three-dimensional structures: characterization and use of generalized valence angle screens, *J. Mol. Graph., 9*, 203–217 (1991).

[5] Shannon, C. E., and Weaver, W. *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, USA, 1963.

[6] Kullback, S. *Information Theory and Statistics*, Dover Publications, Mineola, New York, USA, 1997.

[7] Godden, J. W, Stahura, F. L., and Bajorath, J. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations, *J. Chem. Inf. Comput. Sci., 40*, 796–800 (2000).

[8] Godden, J. W., and Bajorath, J. Differential Shannon entropy as a sensitive measure of differences in database variability of molecular descriptors, *J. Chem. Inf. Comput. Sci., 41*, 1060–1066 (2001).

[9] Feller, W. *An Introduction to Probability Theory and its Applications*, Volume 1, Wiley & Sons Inc., New York, USA, 1950.

[10] Labute, P. Binary QSAR: a new method for the determination of quantitative structure activity relationships, *Pac. Symp. Biocomput., 7*, 444–455 (1999).

[11] McGregor, M. J., and Pallai, P. V. Clustering of large databases of compounds: using MDL "keys" as structural descriptors, *J. Chem. Inf. Comput. Sci., 37*, 443–448 (1997).

[12] Labute, P. A widely applicable set of descriptors, *J. Mol. Graph. Model., 18*, 464–477 (2000).

[13] Stahura, F. L., Godden, J. W., Xue, L., and Bajorath, J. Distinguishing between natural products and synthetic molecules by Shannon descriptor entropy analysis and binary

QSAR calculations, *J. Chem. Inf. Comput. Sci., 40*, 1245–1252 (2000).
[14] Godden, J. W., and 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., 42*, 87–93 (2002).
[15] Henkel, T., Brunne, R. M., Müller, H., and Reichel, F. Statistical investigation into the structural complementarity of natural products and synthetic compounds, *Angew. Chem. Int. Ed. Engl., 38*, 643–647 (1999).
[16] Lee, M.-L., and Schneider, G. Scaffold architecture and pharmacophoric properties of natural products and trade drugs: application in the design of natural product-based combinatorial libraries, *J. Comb. Chem., 3*, 284–289 (2001).
[17] Taskinen, J. Prediction of aqueous solubility in drug design, *Curr. Opin. Drug Discov. Develop., 3*, 102–107 (2000).
[18] Mitchell, B. E., and Jurs, P. C. Prediction of aqueous solubility of organic compounds from molecular structure, *J. Chem. Inf. Comput. Sci*., *38*, 489–496 (1998).
[19] Huuskonen, J., Salo, M., and Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling, *J. Chem. Inf. Comput. Sci., 38*, 450–456 (1998).
[20] Klopman, G., and Zhao, H. Estimation of aqueous solubility of organic molecules by the group contribution approach, *J. Chem. Inf. Comput. Sci., 41*, 439–445 (2001).
[21] Stahura, F. L., Godden, J. W., and Bajorath, J. Differential Shannon entropy analysis identifies molecular descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations, *J. Chem. Inf. Comput. Sci., 42*, 550–558 (2002).
[22] Oprea, T. Property distribution of drug-related chemical databases, *J. Comput.-Aided Mol. Des., 14*, 251–264 (2000).
[23] Sheridan, R. P. The most common chemical replacements in drug-like compounds, *J. Chem. Inf. Comput. Sci., 42*, 103–108 (2002).
[24] Stahura, F. L., Xue, L., Godden, J. W., and Bajorath, J. Design of array-type compound libraries that combine information from natural products and synthetic molecules, *J. Mol. Model., 6*, 550–562 (2000).
[25] Godden J. W., Xue, L., Kitchen, D. B., Stahura, F. L., Schermerhorn, E. J., and Bajorath, J. Median partitioning: a novel method for the selection of representative subsets from large compound pools, *J. Chem. Inf. Comput. Sci., 42*, 885–893 (2002).
[26] Wildman, S. A., and Crippen, G. M. Prediction of physiochemical parameters by atomic contributions, *J. Chem. Inf. Comput. Sci., 39*, 868–873 (1999).
[27] Balaban, A. T. Five new topological indices for the branching of tree-like graphs, *Theor. Chim. Acta, 53*, 355–375 (1979).
[28] Balaban, A. T. Highly discriminating distance-based topological index, *Chem. Phys. Lett., 89*, 399–404 (1982).
[29] Gasteiger, J., and Marsili, M. Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges, *Tetrahedron, 36*, 3219–3228 (1980).

Received on 15 October 2002; Accepted on 2 December 2002