# Identification of Descriptors Capturing Compound Class-Specific Features by Mutual Information Analysis

Anne Mai Wassermann,[†] Britta Nisius,[†] Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received August 20, 2010

The identification of molecular descriptors that contain compound class-specific information is of high relevance in chemoinformatics. A generally applicable way to identify such descriptors is to determine and compare their information content in a given compound activity class and in large databases where the vast majority of compounds do not have the desired activity. For this purpose, the Shannon entropy concept from information theory can in principle be employed. However, previous adaptations of this concept for descriptor profiling are insufficient to select discriminatory descriptors for data sets that dramatically differ in size. Therefore, we introduce a methodology to reliably select such descriptors by transforming the previously introduced differential Shannon entropy formalism into mutual information analysis, another concept from information theory. The newly introduced approach is evaluated by descriptor ranking and correlation analysis on 168 compound activity classes.

## 1. INTRODUCTION

Numerical descriptors of chemical structure and properties play a central role in chemoinformatics, and literally thousands of different descriptors are currently available.[1] Molecular descriptors are often classified as one-, two-, or three-dimensional, depending on the molecular representation from which they are calculated.[2] Descriptors that capture compound class-specific and biological activity-relevant information are of high interest for the exploration of structure−activity relationships. However, the identification of such descriptors is far from being a trivial task. Thus, feature selection approaches are highly desired that are capable of finding descriptors that contain compound class-specific information.[3]

Systematic descriptor selection is complicated by the fact that different descriptors usually have different units and value ranges. Therefore, a direct comparison of molecular descriptors is often difficult, and methods are required that can compare the descriptors and the information they contain regardless of their value ranges. In the context of quantitative structure−activity relationship (QSAR) analysis and database profiling, descriptor selection approaches utilizing the Shannon entropy (SE) concept[4] have previously been developed that meet these basic requirements. The SE concept from information theory provides a basis for the quantification of the information content of data distributions that can be represented as histograms.[5] Thus, SE calculations make it possible to quantify the information content of individual descriptors. For example, it was previously shown that the SE approach can be utilized to compare the variability of various molecular descriptors in different compound databases.[5,6] In order to quantify differences in information content of individual descriptors between different compound data sets, an extension of the SE concept was introduced, the differential SE (DSE) formalism.[7] For example, DSE calculations were carried out to select descriptors distinguishing drug-like molecules from natural products and synthetic molecules.[7] Furthermore, DSE-selected descriptors were successfully utilized to develop binary QSAR models for the prediction of aqueous solubility.[8] In these DSE applications, data sets for descriptor comparison were always of comparable size. However, the identification of descriptors that contain compound class-specific information requires the comparison of data sets that dramatically differ in size, i.e., an activity class containing tens or hundreds of active compounds and a screening database with hundreds of thousands small molecules. We demonstrate that the DSE approach is intrinsically limited in its ability to select information-rich descriptors by comparing compound data sets of very different size. However, by transforming the DSE formalism into mutual information analysis, another concept from information theory, these difficulties can be circumvented. Herein, we introduce the mutual information-DSE (MI-DSE) method that makes it possible to select descriptors that capture activity class-relevant information in an unbiased manner.

## 2. METHODOLOGY

First we describe the SE concept and DSE, its previous extension for descriptor profiling in databases, and then introduce the MI-DSE method for identifying descriptors containing compound class-specific information.

**2.1. SE.** Shannon entropy is a concept from information theory introduced in a seminal paper by Claude Shannon.[4] Originally developed for applications in digital communication, it quantifies the information contained in a "message" distributed over different channels. For molecular descriptor

* Corresponding author. E-mail: bajorath@bit.uni-bonn.de. Telephone: +49-228-2699-306.

[†] These authors contibuted equally to this work.

10.1021/ci100319n 
Published on Web 10/20/2010

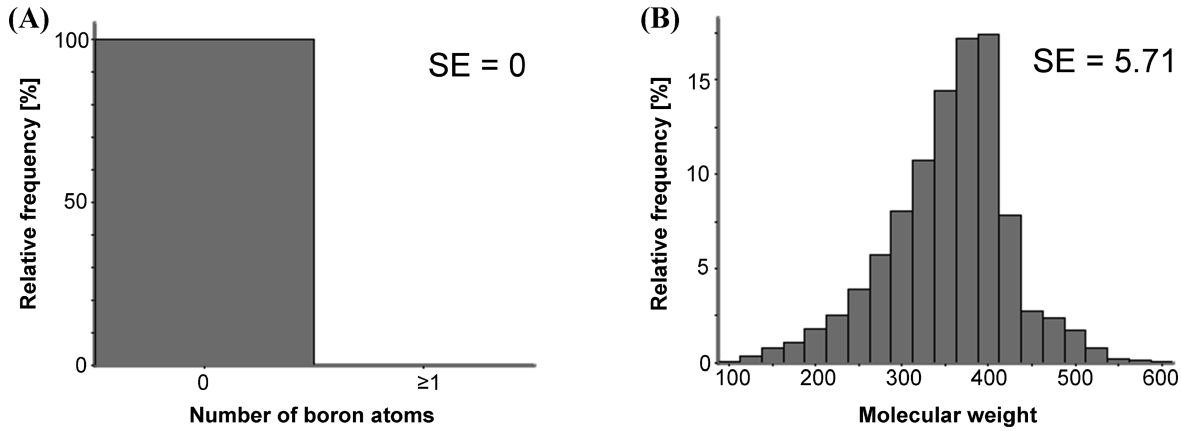


**Figure 1.** Descriptor histograms and corresponding SEs. Exemplary descriptor histograms calculated from 100 000 compounds randomly taken from ZINC are shown, and the corresponding SE values are reported. An exemplary value distribution for a low- and high-entropy descriptor is shown in (A) and (B), respectively.

analysis, this "message" is simply the value of a descriptor calculated for a test compound. The SE concept is applicable to both continuous and discrete value distributions. For descriptor analysis, the focus is on discrete subsets of possible descriptor values, because such subsets can be obtained from any value distributions by discretization of values through a defined binning scheme. Discretization is often facilitated by applying equidistant binning, which partitions a complete value range into a constant number of equally sized subranges.

The SE of a descriptor is given by the average information content of all possible values of the descriptor. The information content of a certain descriptor value depends on the frequency with which this value occurs in a set of compounds. It is provided by the negative of the base 2 logarithm of the frequency of its occurrence $p_i$. Hence, the underlying idea is that a rare descriptor value conveys more specific information about a test compound than a frequently occurring value.

The SE of a descriptor with $n$ possible values is defined as

$$\mathrm{H}(D) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (1)$$

Exemplary value distributions of two different descriptors represented as histograms and the corresponding SE values are shown in Figure 1. None of 100 000 randomly selected ZINC[9] compounds contains a boron atom. Consequently, the corresponding histogram for the descriptor "number of boron atoms" consists of only one bin, and the resulting SE is zero (Figure 1A). By contrast, the values for the descriptor "molecular weight" vary greatly among ZINC database compounds, and therefore this descriptor value distribution yields a high SE reflecting high information content (Figure 1B).

**2.2. DSE.** In order to compare descriptor value settings for any two classes of compounds, e.g., two different databases or active versus inactive compounds, the SE concepts needs to be extended. If we consider the case that value distributions for a given descriptor are very similar for two classes of compounds, i.e., each value occurs with roughly the same frequency in each class, we conclude that this descriptor does not contain class-specific information. By contrast, if the value distributions significantly differ between the two classes, then the descriptor contains such information. However, a descriptor with high SE values (and hence high information content) for both classes might still be nondiscriminatory, if the value distributions are similar. Discriminatory and nondiscriminatory descriptors for comparison of an exemplary activity class and the ZINC database are shown in Figure 2. Most active compounds are positively charged, whereas ZINC molecules are predominantly uncharged or negatively charged (Figure 2A). Thus, the descriptor "formal charge" can be utilized to discriminate between these compound classes. Of course, the discrimination is not perfect because the two histograms overlap. Furthermore, more than 95% of all compounds in both classes do not contain a triple bond. Therefore, histograms for the descriptor "number of triple bonds" are comparable for both data sets, and this descriptor is clearly nondiscriminatory (Figure 2B).

The DSE method was introduced to quantify how much information about a given compound class is contained in the value distribution of a descriptor when compared to another.[7] As illustrated in Figure 3, DSE calculations for a descriptor $D$ and two compound classes **A** and **B** involve the following steps: First, descriptor values are discretized according to an equidistant binning scheme, and histograms are calculated for the two classes **A** and **B**. From these two histograms, the class-specific SEs $\mathrm{H_A}(D)$ and $\mathrm{H_B}(D)$ are calculated. In a next step, the two classes **A** and **B** are combined into a single histogram where the frequencies for a bin $i$ are calculated according to the following equation:

$$f_{\mathbf{AB}}(i) = \frac{n \cdot f_{\mathbf{A}}(i) + m \cdot f_{\mathbf{B}}(i)}{n + m} \qquad (2)$$

Here, $n$ corresponds to the number of molecules in class **A**, $m$ corresponds to the number of molecules in class **B,** and $f_{\mathbf{A}}(i)$ and $f_{\mathbf{B}}(i)$ report bin frequencies for classes **A** and **B**, respectively. The histogram is used to calculate $\mathrm{H_{AB}}(D)$ of the combined classes. Finally, DSE is calculated as

$$\mathrm{DSE}(D) = \mathrm{H_{AB}}(D) - \frac{(\mathrm{H_A}(D) + \mathrm{H_B}(D))}{2} \qquad (3)$$

If the distributions of classes **A** and **B** are similar for a given descriptor, as shown for hypothetical compound classes

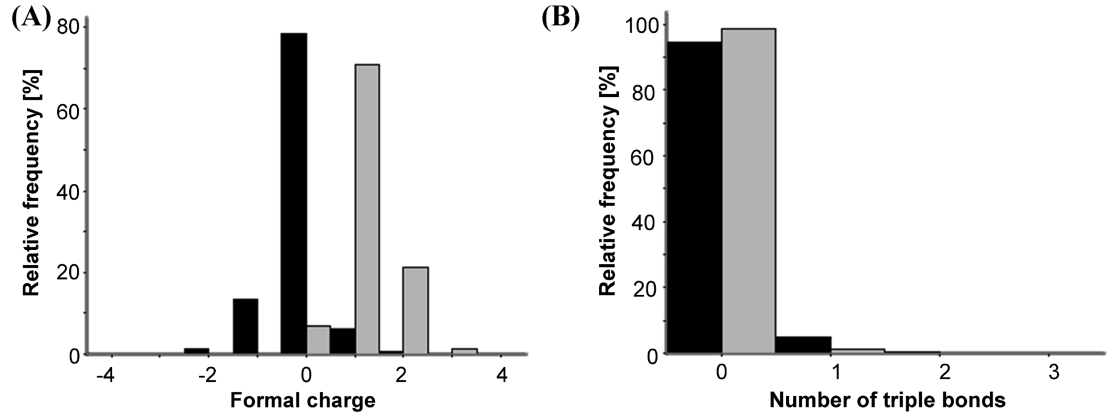


**Figure 2.** Discriminatory and nondiscriminatory descriptors. Exemplary descriptor value distributions are shown for an activity class (gray) and the ZINC compounds (black). (A) Histograms for the "formal charge" descriptor are distinct, and hence this descriptor contains class-specific information. (B) Histograms for "number of triple bonds" are shown that are very similar. Therefore, this descriptor cannot discriminate between the activity class and the reference database.

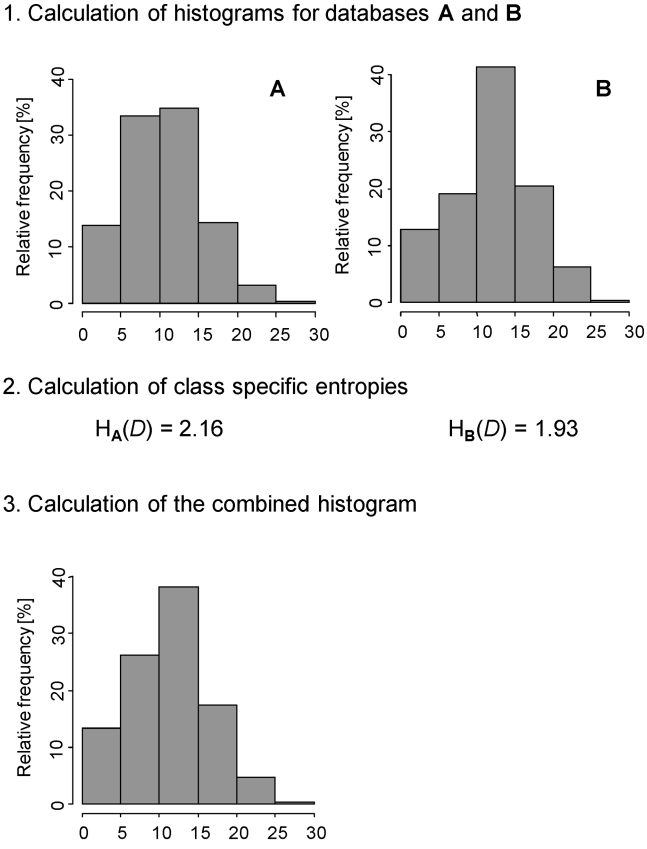


**Figure 3.** DSE calculation. All steps involved in DSE calculation are illustrated for two hypothetical classes of the same size, classes **A** and **B**. In this example, the value range of descriptor *D* is divided into six bins.

of the same size in Figure 3, then $H_{AB}(D)$, $H_A(D)$, and $H_B(D)$ will be similar, and the DSE value will be very small. However, if the value distributions for the two data sets are "complementary", because they have different "shapes" and cover different descriptor value ranges, then $H_{AB}(D)$ of the combined histogram will be larger than the single-class $H_A(D)$ and $H_B(D)$, and consequently, the DSE value will be higher. Hence, DSE values provide an intuitive ranking scheme for descriptors according to their power to discriminate between two different compound classes.

However, a problem arises if the two compound classes are of significantly different size. In this case, the combined histogram is much influenced by the larger class, and its value distribution is biased, as illustrated in Figure 4A. Although the descriptor "number of basic atoms" shows distinct value distributions for an exemplary activity class and ZINC, the combined histogram reflects the descriptor distribution of ZINC. Hence, $H_{AB}(D)$ is essentially equal to $H_B(D)$, and eq 3 can be reduced to

$$\mathrm{DSE}(D) \approx \frac{H_B(D) - H_A(D)}{2} \tag{4}$$

Thus, under these conditions, the magnitude of DSE is mostly determined by descriptors that display high variability in the large compound class but only little variability in the activity class. Then, DSE does no longer quantitatively account for the *value range dependence* of data distributions, which is a key feature of the DSE approach, and the method cannot be applied in a meaningful way.

**2.3. MI-DSE.** When trying to identify descriptors that contain activity class-specific information, we are always faced with large or very large differences in compound class size. Here, the small class represents an activity class, and the large class represents a database where the vast majority of the compounds do not belong to the activity class. Therefore, we have developed a descriptor selection approach that is independent of the size of the compound classes under consideration.

In information theory, the concept that quantifies the amount of information about a class of objects (compounds) captured by a descriptor is known as (average) "mutual information" (MI).[10] MI exactly describes how much information about the class is contained in the value of a descriptor. Formally, MI is defined as the difference between the SE of the descriptor for two combined classes and the conditional SE of the descriptor given the class:

$$\mathrm{MI}(D, C) = H(D) - H(D|C) \tag{5}$$

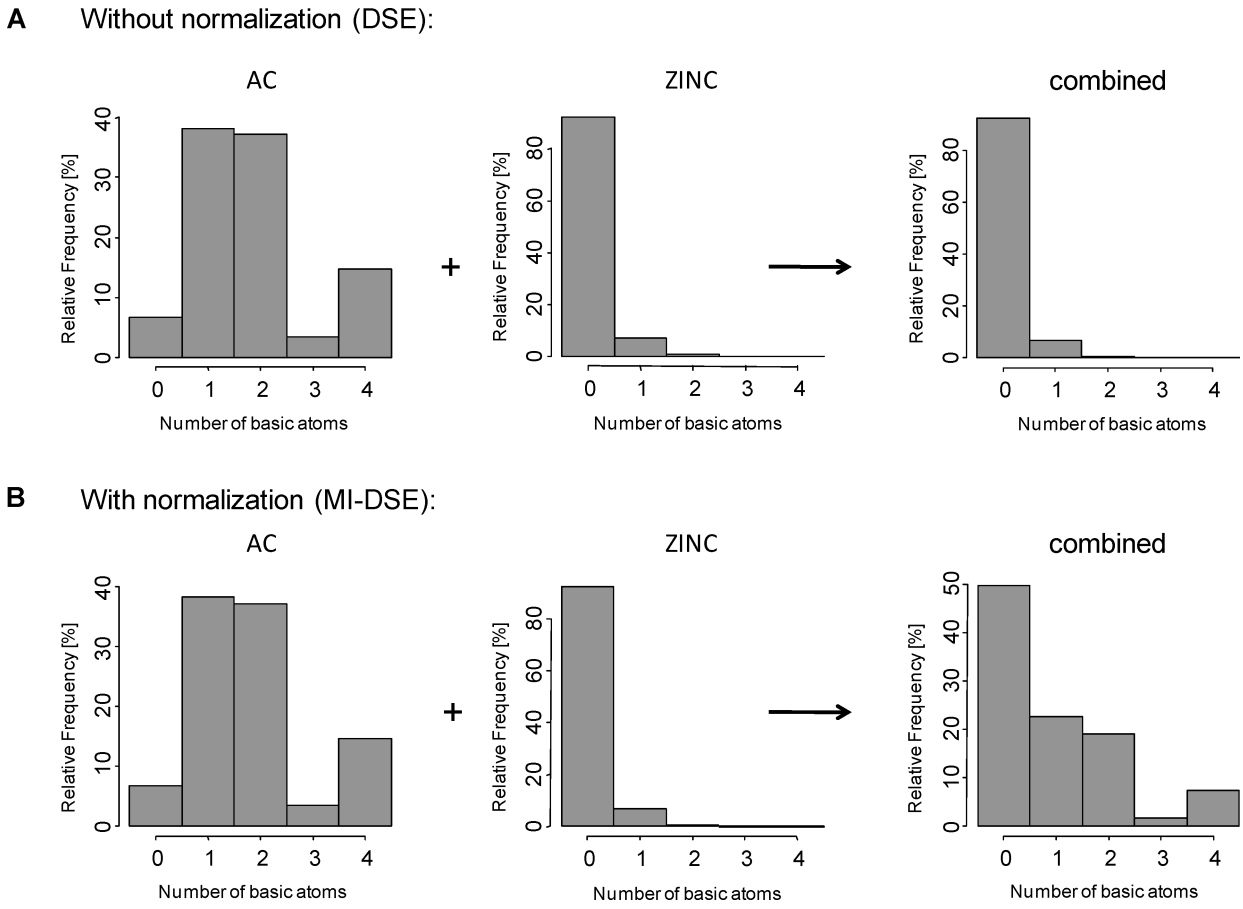


**Figure 4.** Combined histograms for DSE and MI-DSE. Histograms for value distributions of the descriptor "number of basic atoms" are shown for an exemplary activity class, AC, and ZINC. The combined DSE and MI-DSE histograms are shown in (A) and (B), respectively.

Here $D$ is the descriptor and $C$ is the class. H($D$|$C$) quantifies the additional information content of $D$ for class $C$. For two classes **A** and **B**, H($D$|$C$) is given by

$$\mathrm{H}(D|C) = \mathrm{H}_{\mathbf{A}}(D)\cdot\Pr(C = \mathbf{A}) + \mathrm{H}_{\mathbf{B}}(D)\cdot\Pr(C = \mathbf{B}) \tag{6}$$

By setting the probabilities $\Pr(C = \mathbf{A}) = \Pr(C = \mathbf{B}) = 0.5$, we obtain

$$\mathrm{MI}(D, C) = \mathrm{H}(D) - \frac{(\mathrm{H}_{\mathbf{A}}(D) - \mathrm{H}_{\mathbf{B}}(D))}{2} \tag{7}$$

This equation can be seen as an unbiased estimator for the probability that a molecule belongs to either class. Because of the inequality MI($D$,$C$) $\leq$ H($D$) $= 1$, the score is normalized to the range of 0−1.

We now return to the DSE formalism and the calculation of initial value frequencies. Instead of using eq 2 where compound classes were weighted according to their size, we calculate the frequencies as follows

$$f_{\mathbf{AB}}(i) = \frac{f_{\mathbf{A}}(i) + f_{\mathbf{B}}(i)}{2} \tag{8}$$

This eliminates the class size-depending weighting scheme from frequency calculations. On the basis of these frequencies, we can generate the combined histogram of the value distributions of our two compound classes. In the following, we use the term *normalized* for a combined histogram that is based on frequencies calculated according to eq 8 instead of eq 2. The normalized histogram for the descriptor "number of basic atoms" is shown in Figure 4B. In contrast to the original histogram in Figure 4A, the normalized histogram reflects both the value distribution of the descriptor within the activity class and the screening database.

Calculating $\mathrm{H}_{\mathrm{norm}}(D)$ from the normalized histograms yields a modified DSE score that exactly corresponds to eq 7 and is therefore termed MI-DSE:

$$\text{MI-DSE(D)} = \mathrm{H}_{\mathrm{norm}}(D) - \frac{\mathrm{H}_{\mathbf{A}}(D) - \mathrm{H}_{\mathbf{B}}(D)}{2} \tag{9}$$

This quantity also corresponds to the Jensen−Shannon divergence[11] of two feature distributions. The MI-DSE measure has the desired property of yielding normalized scores between 0 and 1, reflecting the significance of descriptors to capture differential information content. A score of 0 indicates that the descriptor distributions for compound classes **A** and **B** are identical and that the descriptor captures no class-specific information, whereas a score of 1 indicates that the value distributions are nonoverlapping and that the descriptor can thus perfectly distinguish between **A** and **B**.

## 3. APPLICATIONS

**3.1. Descriptor Ranking.** To investigate whether MI-DSE and DSE prioritize different descriptors, as would be expected on the basis of our analysis, MI-DSE and DSE calculations were carried out for 168 compound activity

**Table 1.** Compound Activity Classes for Descriptor Comparison[a]

| target | abbreviation | no. ligands |
|---|---|---|
| carbonic anhydrase 2 | CA | 159 |
| endothelin A receptor | EDN | 32 |
| estrone sulfatase | ESU | 35 |
| gonadotropin releasing hormone | GRH | 100 |
| low-density lipoprotein receptor | LDL | 30 |
| lipoxygenase | LIP | 41 |
| squalene synthase | SQS | 42 |
| thromboxane receptor | THR | 33 |
| VEGFR-2 tyrosine kinase | VEG | 36 |
| xanthine oxidase | XAN | 35 |

[a] For each target, its abbreviation and the number of ligands (no. ligands) are reported.

classes extracted from the ChEMBL[12] database that contained at least 50 inhibitors with a minimum potency of 1 $\mu$M. Compounds were represented by a large set of numerical 1D or 2D descriptors available in the Molecular Operating Environment,[13] listed in Supporting Information Table S1. These descriptors accounted for a number of diverse properties, including physicochemical and bulk parameters, atom and bond counts, chemical composition, and surface, topological, or shape properties. This rather comprehensive 2D descriptor set was arbitrarily chosen for our statistical analysis. Other types of descriptors could have also been selected, for example, 3D descriptors. Our analysis aimed at descriptor profiling and ranking to assess the MI-DSE approach and not at finding the best molecular representations to account for specific biological activities. Different binning schemes were investigated by dividing the value ranges of all descriptors into 8, 16, 32, or 64 equally sized bins. For each activity class, all descriptor value distributions were compared to those in a database of 100 000 randomly collected ZINC compounds, MI-DSE and DSE scores were calculated, and the descriptors were ranked in the order of decreasing scores.

In order to compare DSE- and MI-DSE-based rankings for an activity class, Spearman correlation coefficients were calculated. The Spearman correlation coefficient provides a measure for the correlation between two data rankings when the values themselves are not of interest, but the relative order they produce. It can be calculated as the Pearson correlation coefficient of corresponding ranking positions. As a control for the choice of binning schemes, DSE- and MI-DSE-based rankings were first compared among themselves (Supporting Information Table S2). The rankings produced with different numbers of bins were highly correlated for the individual methods, reflecting that descriptor ranking was essentially independent of the utilized number of bins.

We then compared the DSE and MI-DSE rankings and found that, irrespective of the applied binning scheme, correlations between rankings were in most cases not detectable or rather low. The average Spearman rank correlation coefficients were 0.151, 0.201, 0.262, and 0.341 for 8, 16, 32, and 64 bins, respectively (hence, with increasing numbers of bins, the rankings became only slightly more similar). This large-scale comparison over many different compound activity classes demonstrated that DSE and MI-DSE calculations produced very different descriptor rankings.

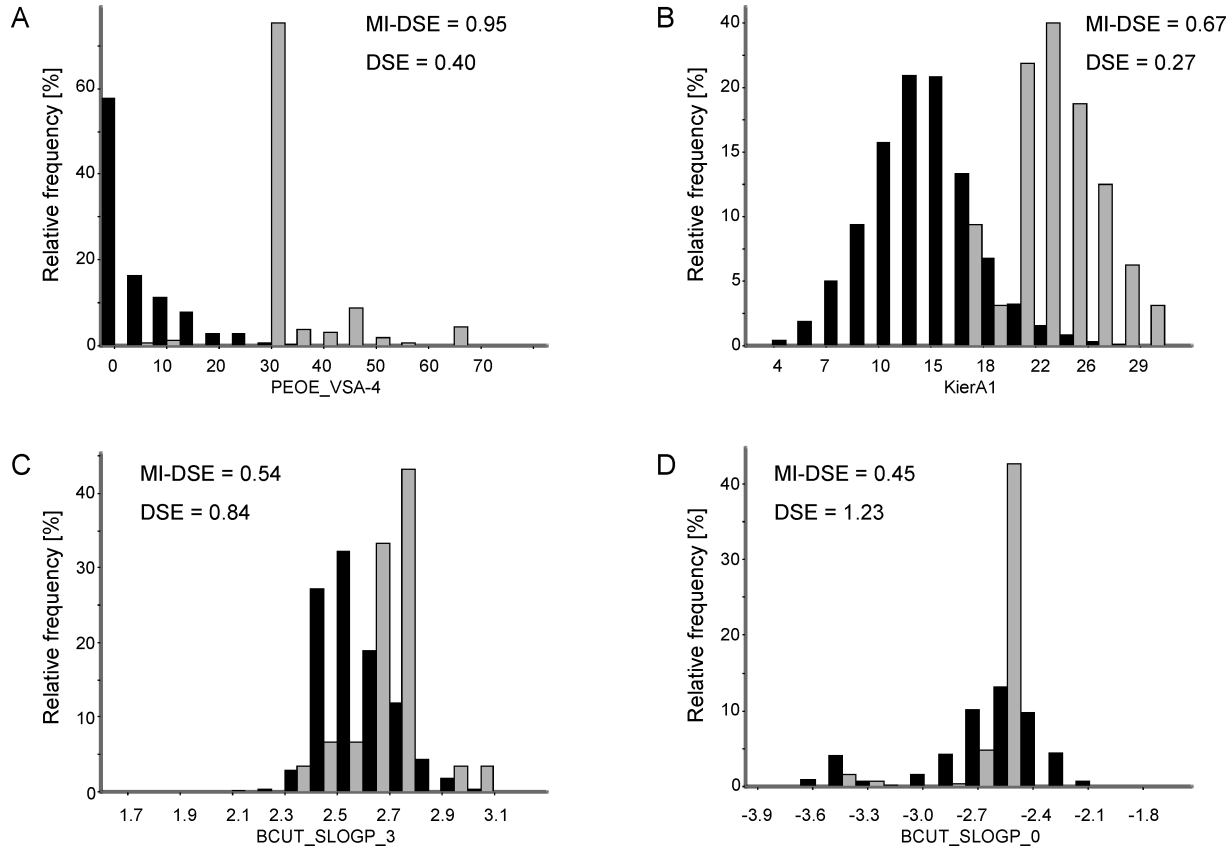


**Figure 5.** Value distributions for top-ranked DSE and/or MI-DSE descriptors. Descriptor distributions for exemplary descriptors for active and database compounds are shown, and the corresponding MI-DSE and DSE values are reported. Descriptor histograms for active and database compounds are shown in gray and black, respectively.

**3.2. Comparison of Top-Ranked DSE and MI-DSE Descriptors.** We also calculated DSE and MI-DSE values for the complete set of descriptors and for 10 other previously reported compound activity classes[14] that contained between 30 and 159 compounds (Table 1) relative to the 100 000 ZINC database compounds. For all descriptors, value ranges were divided into 16 equally sized bins. These 10 activity classes were not included in the descriptor correlation analysis discussed above. For each class, the top-ranked DSE and MI-DSE descriptors were compared (Supporting Information Table S3). The comparison of the descriptor value histograms for active and database compounds and the resulting DSE and MI-DSE values clearly showed that MI-DSE calculations prioritized descriptors capturing compound class-specific information, much more than DSE calculations. Representative examples are shown in Figure 5 (for descriptor definitions, see Supporting Information Table S1). The PEOE_VSA-4 descriptor in Figure 5A is the top-ranked MI-DSE descriptor for activity class CA (class abbreviations are used according to Table 1). This descriptor has a low SE for the ZINC database ($H_B(D) = 1.97$), resulting in a low DSE value. However, the descriptor distributions in active and database compounds display only very little overlap, which results in a high MI-DSE value reflecting discriminatory power of the descriptor. This example illustrates how DSE calculations are dominated by the value distribution of the large compound class. Furthermore, the KierA1 descriptor in Figure 5B is highly ranked for activity class EDN on the basis of MI-DSE calculations but not DSE calculations. This descriptor yields large SE values for both database ($H_B(D) = 3.36$) and active compounds ($H_A(D) = 2.81$), resulting in a low DSE value. However, the MI-DSE value for this descriptor is high because the high-entropy descriptor value distributions for both sets overlap only a little, revealing that the descriptor captures compound class-specific information. In Figure 5C, value distributions for BCUT_SLOGP_3 are shown that is the top-ranked descriptor for activity class LDL on the basis of both DSE and MI-DSE analysis. This descriptor produces a high SE value for the database ($H_B(D) = 3.56$) and a low SE value for the activity class ($H_A(D) = 1.92$), resulting in the high DSE value, which is determined by the high SE value for the database. However, the value distributions of this descriptor display only limited overlap, which also results in a high MI-DSE value, consistent with its evident discriminatory nature. Finally, in Figure 5D, distributions are shown for BCUT_SLOGP_0, the top-ranked DSE descriptor for activity class CA. This descriptor has a high SE value for database ($H_B(D) = 3.30$) but a low SE value for active compounds ($H_A(D) = 0.83$), because the majority of their descriptor values populate a single bin. These value distributions thus result in a high DSE value, although the descriptor does not contain class-specific information. This is the case because the peaks of the data distributions closely overlap, i.e., most active and database compounds produce similar values. This is clearly indicated by the low MI-DSE value of this descriptor and its low MI-DSE rank.

## 4. CONCLUSIONS

The identification of descriptors that capture compound class-specific information is of high relevance for many chemoinformatics applications. Generally, compound class-specific information must be assessed by comparing sets of compounds having a desired property (such as, for example, biological activity) with data sets where all or at least the majority of compounds lack this property. The larger this reference database is, the more reliable the assessment of class-specific information becomes. If only small reference sets are utilized, then the analysis is not meaningful. For the study of biological activity, descriptors that systematically differ in their value settings between active and database compounds are highly desired, i.e., descriptors that are activity relevant. Here we have introduced an information theoretic approach to reliably assess compound class-specific information content of descriptors that is not biased by intrinsic differences in the size of activity classes and reference databases. This has been accomplished by combining the differential Shannon entropy formalism with the mutual information concept. The comparison of value distributions of descriptors and the resulting DSE and MI-DSE values and descriptor rankings has confirmed the utility of the MI-DSE approach to identify descriptors containing class-specific information. This newly introduced approach is straightforward and should be useful for large-scale feature analysis.

**Supporting Information Available:** Tables S1−S3 report descriptors, rank correlation coefficients, and DSE and MI-DSE descriptor rankings, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley: Weinheim, Germany, 2000.
(2) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(3) Liu, Y. A comparative study on feature selection methods for drug discovery. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1823–1828.
(4) Shannon, C. E. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423.
(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) Stahura, F. L.; Godden, J. W.; Bajorath, J. Distinguishing between natural products and synthetic molecules by descriptor Shannon entropy analysis and binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245–1252.
(7) 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.
(8) Stahura, F. L.; Godden, J. W.; Bajorath, J. Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 550–558.
(9) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(10) Cover, T. M.; Thomas, J. A. *Elements of information theory*; Wiley: New York, 1991.
(11) Lin, J. Divergence measures based on Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.
(12) ChemblDB; European Bioinformatics Institute (EBI): Cambridge, U.K.; http://www.ebi.ac.uk/chembl/. Accessed August 15, 2010.
(13) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2007.
(14) Nisius, B.; Vogt, M.; Bajorath, J. Development of a fingerprint reduction approach for Bayesian similarity searching based on Kullback-Leibler divergence analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.

CI100319N