Variability of Molecular Descriptors in Compound Databases Revealed by Shannon Entropy Calculations

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

Computational Chemistry and Informatics, New Chemical Entities, 18804 North Creek Parkway, Bothell, Washington 98011

Received January 5, 2000

A method is introduced to calculate and compare the variability of molecular descriptors in compound databases. Descriptor variability analysis is based on histograms recording the distribution of molecular descriptors and calculation of Shannon entropy (SE), a metric originally applied in digital communication. SE values reflect the variability of descriptor settings. We have calculated a total of 92 molecular descriptors in the ACD and NCI databases and ranked them according to their variability. Significant differences in entropy are observed for a number of descriptors. However, the most variable descriptors are similar in the ACD and NCI databases. Such high-entropy descriptors are preferred tools to discriminate between compounds or account for the diversity of chemical libraries.

INTRODUCTION

Molecular descriptors typically account for physicochemical properties, molecular topology/connectivity, conformational parameters, or structural fragments. Descriptor-based representations of molecules are widely used to analyze molecular similarity, describe the diversity of libraries and cluster/partition molecules, or study structure−activity relationships and druglike properties of compounds.[1−10] The performance of sets of descriptors for specific tasks such as, for example, classification of compounds according to biological activity or diversity analysis, has been evaluated in a number of case studies.[3−8] We intended to systematically analyze the distribution of molecular descriptors in large compound databases and determine their variability. This would make it possible, for example, to select highly variable descriptors for diversity analysis or other applications. However, this study was complicated by the fact that the distributions of many descriptors cannot be directly translated into variability because their units and value ranges differ. For example, "log *P*" or descriptors accounting for "molecular volume" present a continuum of values, whereas descriptors counting the number of "aromatic bonds" in a molecule typically adopt a narrow range of discrete values. Moreover, a "structural key" (i.e., molecular fragment) is either present or absent in a molecule and thus binary in nature. Therefore, a more general analysis and comparison of descriptor variability requires a uniform data representation that is independent of the type of descriptors and the value ranges they can adopt.

To achieve these ends, we have generated histograms of descriptor distributions using a consistent binning scheme, which permits comparison of these distributions. On the basis of this consistent data representation, descriptor variability was calculated by application of an entropic formulation, termed "Shannon entropy" (SE).[11] This concept was originally applied in digital communication technology to determine the amount of data that could be transmitted within a given range of frequencies. Herein we report our histogram- and SE-based method to capture descriptor variability independent of value ranges and distributions. To apply the method, we have determined and compared the entropy of molecular descriptors in two large compound databases.

METHODS

Shannon entropy is defined as

$$\mathrm{SE} = -\sum p_i \log_2 p_i \qquad (1)$$

In this formulation, p is the probability of a data point computed from a "count" (c) that adopts a value within a specific data interval i. Thus, p is calculated as

$$p_i = c_i / \sum c_i \qquad (2)$$

Equation 1 that converts probability to the Shannon entropy contains a logarithm to the base 2, which corresponds to a scale factor. Probabilities and, in turn, SE values can be calculated for any set of data that is divided into evenly spaced intervals (bins). SE values for different data sets can be directly compared, provided the binning scheme is uniform. This is the case when data sets are represented in histograms where the data range is divided into the same number of bins. As long as the number of data intervals between the minimum and maximum value is held constant, SE values are independent of the size of the interval. Figure 1 illustrates how SE values are calculated from a histogram that counts the number of data points per data interval.

How can this concept be applied to analyze and compare the variability of molecular descriptors? We need to calculate molecular descriptors for all compounds in a database and produce histograms that cover the entire value range of each

* To whom correspondence should be addressed at New Chemical Entities. Phone: (425) 424-7297. Fax: (425) 424-7299. E-mail: jbajorath@nce-mail.com.

† New Chemical Entities and Department of Biological Structure, University of Washington, Seattle, WA 98195.

10.1021/ci000321u CCC: $19.00
Published on Web 04/19/2000

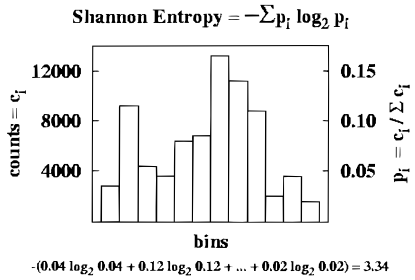

Figure 1. Shannon entropy calculated from a hypothetical descriptor distribution. Data counts falling within one of a fixed number of bins (12 in this case) are accumulated. These counts are converted into sample probabilities by dividing each count by the summed counts of all bins. The resulting probabilities are used to calculate the SE value of 3.34 for this data distribution.

descriptor using the same number of bins. Although molecular descriptors have intrinsically different numbers of possible values (e.g., comparing log *P* and number of aromatic bonds), SE values can be directly compared. For a fixed number of intervals (*i*'s), SE reaches a maximum value when all p_i have the same value, i.e., when all the bins are equally populated. By contrast, if all the occurrences fall within a single bin ($p_i = 1.0$), then SE has the value of zero.

In this study, we analyzed the variability of a total of 92 molecular descriptors that could be calculated from 2D representations of molecules. These diverse descriptors are summarized in Table 1. They include bulk properties, physicochemical parameters, atom or bond counts, topology and shape indices, and structural keys. Descriptors were calculated with MOE[12] for compounds (single molecules, not including any noncovalent complexes) in the ACD[13] and NCI[14] databases. ACD contains many organic compounds that are often used as reagents, while NCI consists of compounds evaluated as potential anti-cancer therapeutics. Descriptors for 199 420 ACD and 122 264 NCI compounds were calculated in this study. Histograms of descriptor distributions were consistently generated using 100 data intervals. This empirically derived number of bins yielded graphically meaningful data dispersion. The magnitude of SE values depends on the number of bins. For 100 bins used here, the maximum entropy value is approximately 6.64 (for even data dispersion over all bins). For comparison, for 50 bins, the theoretical maximum entropy value would be 5.64. Statistical analysis of descriptor distributions and SE calculations were performed with Perl programs written by the authors.

RESULTS AND DISCUSSION

The concept of Shannon entropy is applied to determine the variability of molecular descriptors in large compound databases. Calculation of SE values for direct comparison critically depends on consistent representation of data, which are highly diverse and have very different value ranges. Therefore, we have represented distributions of different descriptors as histograms, where each observed data range was divided into 100 bins of equal size. Figure 2 shows representative histograms of molecular descriptors for ACD and NCI compounds and SE values calculated from these histograms. The comparison confirms that, as to be expected, narrow distributions of descriptor values result in lower descriptor entropy than broad distributions. In our analysis, this correlation was valid for all 92 descriptors studied in histograms, consistent with the idea that SE values are proportional to the amount of choices that are available to a system.[11] For the analysis of descriptor distributions, this can be interpreted as a "nonparametric" estimator of data spread, which is sensitive to the possible values that the descriptor can attain. For example, a "binary" descriptor (e.g., structural

Table 1. Summary of Classes of Molecular Descriptors Analyzed in this Study[a]

	Atom Counts
a−aro	number of aromatic atoms
a−nC	number of carbon atoms
a−nX	number of atoms of element X
a−ICM	entropy of the elemental distribution
	Bond Counts
b−aro	number of aromatic bonds
b−double	number of double nonaromatic bonds
b−rotR	fraction of nonring bonds
b−1rotR	fraction of single nonring bonds
HB-a	number of hydrogen bond acceptors
HB-d	number of hydrogen bond donors
	Connectivity Indices[b]
chi1	sum of inverse square roots of $d_i d_j$ over all bonded heavy atoms
chi1v	sum of inverse square roots of $v_i v_j$ over all bonded heavy atoms
	Graph Distances[c]
VdistMa	$\sum a_{ij} \log a_{ij}/s - \log s$
VdistEq	entropy of the distribution of the distance matrix
$^i\kappa$ (KierN)	Kier κ shape indices
BalabanJ	Balaban connectivity index
	Charge
PEOE−RPC	magnitude of the largest (positive/negative) charge divided by the sum of such charges
PEOE−VSA	sum of the van der Waals surface areas of atoms of designated partial charge range
bpol	sum of the absolute value difference between atomic polarizabilities

[a] Classes of molecular descriptors are shown that are representative of 89 (of the 92) descriptors analyzed in this study. For example, the charge descriptor "PEOE−VSA" has 20 variants. In addition, the set of 92 descriptors includes three bulk properties, molecular weight, refractivity, and log *P*. [b] d_i = number of heavy atoms bonded to atom *i*. p_i = number of atom *i*'s s, p electrons. h_i = number of hydrogens bonded to atom *i*. z_i = atomic number of atom *i*. $v_i = (p_i - h_i)/(z_i - p_i - 1)$. [c] *s* = sum of the distance matrix entries. a_{ij} = the shortest path from atoms *i* and *j*.

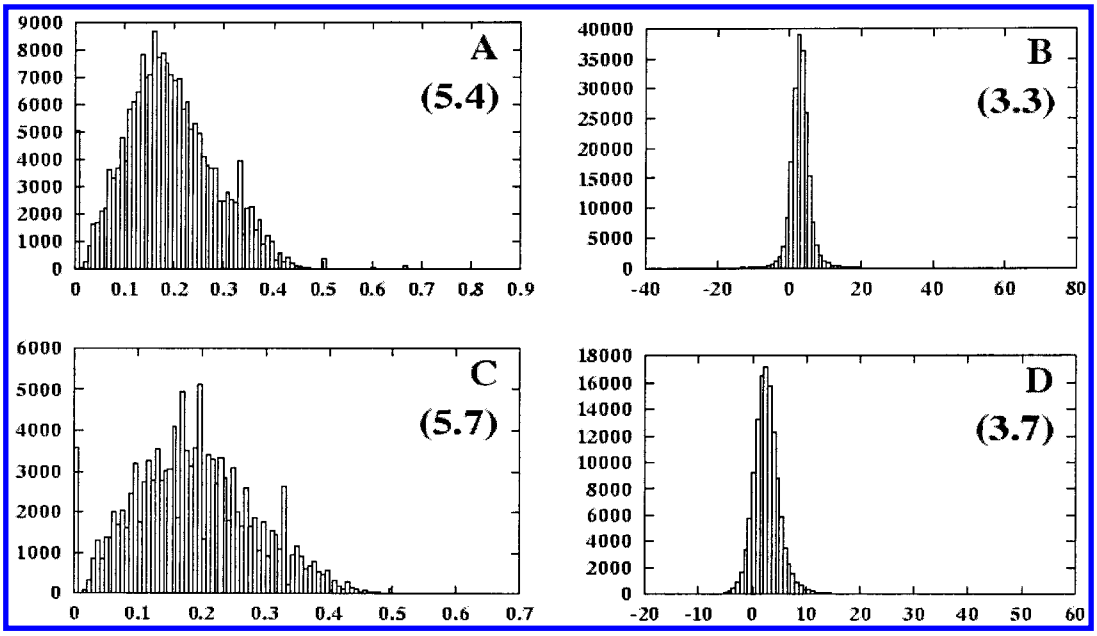

Figure 2. Histograms of descriptor distributions. Each histogram consists of 100 bins covering the entire data range. (A) and (C) report the distribution of descriptor "b−1rotR" (the fraction of single nonring bonds); (B) and (D) are for "log P(o/w)" (the logarithm of the octanol/water partition coefficient). (A) and (B) are from the ACD database; (C) and (D) are from the NCI database. Calculated SE values are given in parentheses.

key) can only assume two values (either 0 or 1), and its maximum entropy is 1.0, if there are equal proportions of both values. By contrast, the entropy of descriptors with continuous value ranges (e.g., "molecular weight") could reach the theoretical maximum of ∼6.64 (for 100 data intervals).

It is important to consider why the variability of descriptors cannot be systematically compared using simple statistic measures such as standard deviations. A standard deviation, or any statistic derived from variance, depends upon the identification of a central mean or mode. However, as to be expected, we found that descriptor distributions are often not Gaussian-like but multimodal, which makes it difficult, if not impossible, to calculate central values. By contrast, calculation of Shannon entropy is independent of such values. More importantly, unlike SE values, standard deviations must be provided in units of the data collected, which prohibits the comparison of the variability of descriptors with different units. SE values depend on consistent data representation but are unit-independent.

In Table 2, the most and least variable molecular descriptors in the ACD database are reported and compared to NCI descriptors. In both databases, a wide range of SE values is observed, from 0 to close to 6 (i.e., approaching the upper SE limit for 100 bins). We would expect that both the intrinsic variability of molecular descriptors and the characteristics of compounds under investigation determine descriptor entropy. Consistent with this view, ranking of descriptors according to their variability in ACD and NCI revealed that the most variable descriptors were similar but not identical in both databases. Descriptors with highest variability account for molecular flexibility, partial charges, and topology indices of compounds. By contrast, a number of molecular descriptors, including simple charge summations, triple bond counts, and counts of halogen atoms, have little, if any, variability. Thus, these descriptors would not be useful to compare compounds in these databases. Since

Table 2. Descriptors with Highest and Lowest Entropy[a]

	ACD		NCI	
descriptor	rank	entropy	rank	entropy
b−1rotR	1	5.417	1	5.703
b−rotR	2	5.368	2	5.647
a−ICM	3	5.368	3	5.312
VadjEq	4	5.310	6	5.149
PEOE−RPC−	5	5.310	4	5.311
VdistEq	6	5.130	5	5.152
VdistMa	7	5.059	8	4.936
PEOE−RPC+	8	4.986	7	4.978
VadjMa	9	4.837	9	4.661
BalabanJ	10	4.608	10	4.434
WeinerPol	11	4.118	16	3.936
Zagreb	12	4.072	19	3.869
a−nF	81	0.740	81	0.338
b−triple	82	0.413	83	0.285
a−nBr	83	0.373	82	0.315
WeinerPath	84	0.247	86	0.154
a−nP	85	0.176	84	0.211
KierFlex	86	0.126	85	0.199
a−nI	87	0.102	87	0.081
Fcharge	88	0.006	89	0.000
RPC+	89	0.000	89	0.000
RPC−	89	0.000	89	0.000
PC+	89	0.000	89	0.000
PC−	89	0.000	89	0.000

[a] The 12 most and least variable descriptors in ACD are shown and compared to their rank in NCI.

ACD contains many reagents commonly used in synthetic and combinatorial chemistry, descriptors with low entropy in ACD may not be suitable for diversity design.[15,16] On the contrary, high-entropy descriptors are expected to perform well in this and related cases, for example, similarity searching.[17,18]

Entropy values calculated for ACD and NCI compounds correlate in an interesting way, as shown in Figure 3. High- and low-entropy descriptors are similar in these databases (see also Table 2). All of these descriptors are summarized

Table 3. Definitions of the Descriptors with High and Low Entropy

a. Physicochemical Properties and Charge Descriptors[a]

descriptor	definition
Fcharge	total charge of the molecule (sum of formal charges)
PC+	sum of the positive q_i
PC−	sum of the negative q_i
RPC+	largest positive q_i divided by the sum of the positive q_i
RPC−	smallest negative q_i divided by the sum of the negative q_i
PEOE−RPC+	largest positive p_i divided by the sum of the positive p_i
PEOE−RPC−	smallest negative p_i divided by the sum of the negative p_i
PEOE−VSA−6	sum of v_i where p_i is less than −0.30
PEOE−VSA−1	sum of v_i where p_i is in the range [−0.10, −0.05)
PEOE−VSA−0	sum of v_i where p_i is in the range [−0.05, 0.00)
PEOE−VSA+0	sum of v_i where p_i is in the range [0.00, 0.05)
PEOE−VSA+1	sum of v_i where p_i is in the range [0.05, 0.10)
bpol	sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule

b. Atom, Bond, and Electron Pair Count Descriptors

descriptor	definition
a−ICM	entropy of the distribution of elements in the molecule
a−nF	number of fluorine atoms
a−nBr	number of bromine atoms
a−nP	number of phosphorus atoms
a−nI	number of iodine atoms
b−1rotR	fraction of single nonring bonds
b−rotR	fraction of nonring bonds
b−triple	number of triple bonds
VadjEq	entropy of the distribution of values in the adjacency matrix
VadjMa	log (2 times the number of bonds)
HB-a	rule-based definition of the number of hydrogen bond acceptors[7]

c. Connectivity Indices, Kier κ Shape Indices,[20] Graph Distance Matrix Descriptors, and Surface Area Descriptors[b]

descriptor	definition
Zagreb	sum of the squares of the d_i of the heavy atoms
Chi1v	sum of the inverse square roots of $v_i v_j$ for all bonded heavy atoms i and j
Kier3	$(n-1)(n-3)^2/p_3^2$ for odd n, and $(n-3)(n-2)^2/p_3^2$ for even n
KierA3	$(s-1)(s-3)^2/p_3^2$ for odd n, and $(s-3)(s-2)^2/p_3^2$ for even n, where $s = n + a$
KierFlex	(KeirA1)(KeirA2)/n
WienerPath	Wiener path number, which is half the sum of all the distance matrix entries[21]
WienerPol	Wiener parity number, which is half the number of entries in the distance matrix with a value of 3 (ref 21)
VdistMa	if m is the sum of the distance matrix entries, then VDistMa is defined as the sum $a_{ij} \log a_{ij}/m - \log m$ over all i and j (logarithms are taken in base 2)
VdistEq	entropy of the distribution of values in the distance matrix
BalabanJ	Balaban's connectivity topological index[22]

[a] q_i is the partial charge of atom i in a molecule, p_i represents the partial charge of atom i calculated according to the PEOE method,[19] and v_i is the van der Waals surface area of atom i. [b] d_i is the number of heavy atoms bonded to atom i. $v_i = (p_i - h_i)/(z_i - p_i - 1)$, where p_i is the number of s and p valence electrons of atom i, h_i is the number of hydrogens bonded to atom i, and z_i is the atomic number of atom i. n is the number of atoms in the non-hydrogen graph of the molecule, m is the number of bonds, and a is the sum of $r_i/r_c - 1$. r_i is the covalent radius of atom i, and r_c is the covalent radius of a carbon atom. p_3 is the number of paths of length 3. The graph distance matrix of a molecule with n atoms is defined as the $n \times n$ matrix, A, where a_{ij} is the length of the shortest path in the graph between atoms i and j. The descriptors represent values derived from the graph distance matrix of the non-hydrogen molecular graph of a molecule.

and defined in Table 3. However, the majority of the descriptors have midrange entropy (i.e., SE values between 2 and 4) in both databases, and in this range, some significant differences in entropy are observed. The "off-diagonal" descriptors in region B of Figure 3 have maximum differences in SE values of 0.5−1.0, including, among others, shape and connectivity indices and hydrogen bond acceptors (Kier3, KierA3, PEOE−VSA, chi1v, HB-a, bpol). These descriptors have higher entropy (and thus greater data spread) in NCI than in ACD and may thus be suitable, for example, to distinguish NCI from ACD compounds. These findings imply that SE analysis of descriptor variability should aid in the identification of descriptors that are capable of capturing compound library- or class-specific features, for example, druglike properties.

CONCLUSIONS

We have aimed at a general analysis of the variability of molecular descriptors in compound databases, irrespective of their units and value ranges, and identified Shannon entropy as a suitable concept. SE calculations can be carried out to compare the variability of very different descriptors and are generally applicable to analyze descriptor distributions in compound databases. Although we have, for practical purposes, limited our analysis to those descriptors that could be calculated from 2D representations of molecules, descriptors that depend on molecular conformation can also be studied using this approach. As a first application, calculations on ACD and NCI compounds were presented and revealed both similarities and differences in the relative variability of descriptors. Differences in descriptor entropy

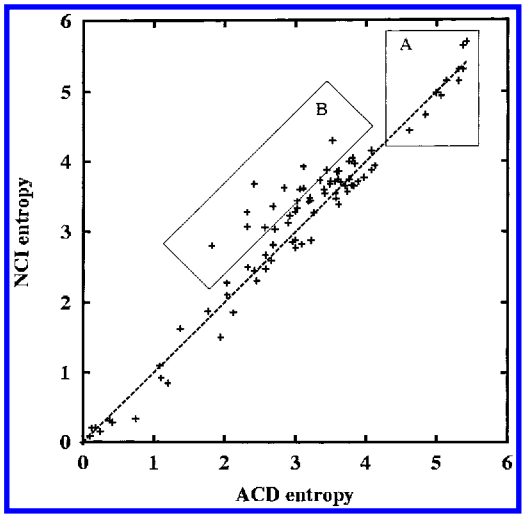

Figure 3. Correlation of descriptor entropy values between ACD and NCI compounds. Descriptors in region A show significant and similar variability in both databases. Descriptors in region B show significant differences in variability in ACD and NCI. Descriptors in region A: PEOE−RPC−, PEOE−RPC+, VadjEq, VadjMa, VdistEq, VdistMa, a−ICM, b−1rotR, b−rotR, and balabanJ. Descriptors in region B: KierA3, Kier3, PEOE−VSA+0, PEOE−VSA−0, PEOE−VSA+1, PEOE−VSA−1, PEOE−VSA−6, chi1v, HB-a, and bpol (for definitions, see Table 3).

provide a possible route to detect class-specific features of compounds. In addition, highly variable descriptors are preferred in the analysis of molecular diversity.

ACKNOWLEDGMENT

We thank Ling Xue for scientific discussions and help in the assembly and calculation of molecular descriptors.

REFERENCES AND NOTES

(1) Brown, R. D. Descriptors for diversity analysis. *Perspect. Drug Discovery Des.* **1997**, *7/8*, 31−49.
(2) Bures, M. G.; Martin, Y. C. Computational methods in molecular diversity and combinatorial chemistry. *Curr. Opin. Chem. Biol.* **1998**, *2*, 376−380.
(3) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D molecular descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 731−740.
(4) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: Using MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(5) Matter, H. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.
(6) Matter, H.; Pötter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211−1225.
(7) Xue, L.; Godden, J.; Gao, H.; Bajorath, J. Identification of a preferred set of molecular descriptors for compound classification based on principal component analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 669−704.
(8) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of large structure/biological activity data sets using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(9) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(10) Ajay; Walters, P.; Murcko, M. A. Can we learn to distinguish between "drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(11) Shannon, C. E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Urbana, IL, 1963.
(12) MOE (Molecular Operating Environment), Chemical Computing Group Inc., 1255 University St., Montreal, Quebec, Canada H3B 3X3.
(13) ACD (Available Chemicals Directory), available from MDL Information Systems Inc., 14600 Catalina St., San Leandro, CA 94577.
(14) NCI (National Cancer Institute) publicly available compound database; see also: Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paull, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.
(15) Martin, E. J.; Blaney, J. M.; Siani, M. A.; Spellmeyer, D. C.; Wong, A. K.; Moos, W. H. Measuring diversity: Experimental design of combinatorial libraries for drug discovery. *J. Med. Chem.* **1995**, *38*, 1431−1436.
(16) Mason, J. S.; Hermsmeier, M. A. Diversity assessment. *Curr. Opin. Chem. Biol.* **1999**, *3*, 342−349.
(17) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(18) Xue, L.; Godden, J.; Bajorath, J. Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 881−886.
(19) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(20) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.
(21) Balaban, A. T. Five new topological indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.
(22) Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.

CI000321U