Mold^2, Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics

Huixiao Hong,*[,†] Qian Xie,[‡] Weigong Ge,[†] Feng Qian,[‡] Hong Fang,[‡] Leming Shi,[†] Zhenqiang Su,[†] Roger Perkins,[‡] and Weida Tong[†]

Center for Toxicoinformatics, Division of Systems Toxicology, National Center for Toxicological Research, and Division of Bioinformatics, ZTech, an ICF International Company at the National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, Arkansas 72079

Received February 4, 2008

Research applications in chemoinformatics and toxicoinformatics increasingly use representations of molecules in the form of numerical descriptors that capture the structural characteristics and properties of molecules. These representations are useful for ADME/toxicity prediction, diversity analysis, library design, QSAR/QSPR, virtual screening, and other purposes. Molecular descriptors have ranged from relatively simple forms calculated from simple two-dimensional (2D) chemical structures to more complex forms representing three-dimensional (3D) chemical structures or complex molecular fingerprints consisting of numerous bit positions to represent specific chemical information. The Mold^2 software was developed to enable the rapid calculation of a large and diverse set of descriptors encoding two-dimensional chemical structure information. Comparative analysis of Mold^2 descriptors with those calculated by Cerius^2, Dragon, and Molconn-Z on several data sets using Shannon entropy analysis demonstrated that Mold^2 descriptors convey a similar amount of information. In addition, using the same classification method, slightly better models were generated using Mold^2 descriptors compared to those generated using descriptors from the compared commercial software packages. The low computing cost for Mold^2 makes it suitable not only for small data sets, such as in QSAR, but also for large databases in virtual screening. High reproducibility and reliability are expected because Mold^2 does not require 3D structures. Mold^2 is freely available to the public (http://www.fda.gov/nctr/science/centers/toxicoinformatics/index.htm).

INTRODUCTION

Computational approaches applied in drug discovery and toxicity prediction often require molecular descriptors that reflect structural information and physicochemical properties of chemicals. Predictive models in quantitative structure−activity relationship (QSAR)/quantitative structure−property relationship (QSPR) are developed by correlating the properties or activities with the structural information. The structures themselves are difficult for use in the models. Therefore, the issue becomes how the structural information is represented in the models. Molecular descriptors are used to extract the structural information in the form of numerical or digital representation that is suitable for model development, serving as the bridge between the molecular structures and physicochemical properties or biological activities of chemicals.

In 1937, Louis L. Hammett developed an equation that became a historic milestone in chemistry. In the Hammett equation, the rate or equilibrium constants of side-chain reactions of aromatic acids, phenols, and anilines as well as other compounds are calculated from two parameters: ρ characterizing the nature of the reaction and σ quantifying the electronic effect of replacing a hydrogen atom by a given substituent in the *meta* or *para* position.[1] The mechanistically based Hammett equation thus became the first QSPR equation in chemistry.

The QSAR paradigm has its roots in the early 1960s when Hansch[2] and Fujita,[3] building on the work of Hammett[1] and Taft,[4] introduced a steric parameter π, the relative hydrophobicity of a constituent. The hydrophobicity parameter π based on the *n*-octanol/water partition coefficients enabled the extension of QSPR to the realm of biology (QSAR) and has since been extensively used to develop quantitative models from *in vitro* activity data.

The parameters σ and π in the classic Hansch equation are chemical properties that in the QSAR lexicon are one-dimensional (1D) molecular descriptors that were initially empirically determined but later on were often computationally calculated. Fostered by the power of computing, extensive interests and broadly diverse approaches in the last two decades focused on developing molecular descriptors and their correlation with biological activity. 1D descriptors remain important and are often used with, and sometimes instead of, descriptors derived from both two-dimensional (2D) and three-dimensional (3D) structures.[5–39] Concomitantly, programs became available for computing descriptors either as independent software or as part of QSAR software; notable examples are ADAT,[40] CODESSA,[41] OASIS,[42] POLLY,[43] CERIUS^2 (www.accelrys.com), DRAGON (www.

* Corresponding author phone: (870)543-7296; fax: (870)543-7382; e-mail: huixiao.hong@fda.hhs.gov. Corresponding author address: Center for Toxicoinformatics, Division of Systems Toxicology, NCTR, 3900 NCTR Road, Jefferson, AR 72079.

† Center for Toxicoinformatics, Division of Systems Toxicology, National Center for Toxicological Research.

‡ Division of Bioinformatics, ZTech.

10.1021/ci800038f
Published on Web 06/20/2008

talete.mi.it), MOLCONN-Z (www.eslc.vabiotech.com), SYBYL (www.tripos.com), and others.

In the most general sense, an effective molecular descriptor is the one representing chemical structure features or chemical properties that vary across a set of chemicals similarly to a biological end point associated with the chemicals (i.e., the descriptor and the biological end point should be correlated). Early research emphasized descriptors amenable to physical interpretation in a biological context which, in turn, could provide insights of an underlying mechanism of action, ways of reducing a drug's toxicity or increasing its efficacy. Time and experience, however, taught researchers that the physically interpretable descriptors, as desirable as they were, did not usually yield the most accurate QSAR model. Rather, regression and classification models derived from very large descriptor sets of esoteric physical meaning often yielded models with greater accuracy and fidelity.

Over time, the descriptor software evolved to produce hundreds or, in the case of 3D CoMFA, thousands of descriptors. While using more descriptors increased the fidelity of structure−activity correlation, it resulted in a new problem: the overfitted models. With descriptors usually outnumbering the chemicals available for training, a model with an excellent fit to the training set could be useless for predicting unknown chemicals if the model is overfitted. The specter of an overfitted model demanded both diligence in descriptor selection and vigilance in guarding against overfitted models, both of which add an immense computational expense to model development. The Mold^2 descriptors reported here are large in number, warranting vigilance when many are used. In drug discovery where high throughput technology and combinatorial chemistry yield huge databases of biological activities and chemical structures, the computational expense is amplified many times over.

While highly desirable, the *a priori* design and validation of a relatively smaller number of simpler molecular descriptors that can give a comparable or a better result than a large descriptor set for drug candidate discovery and toxicity prediction is a formidable task. Caution must be exercised to ensure that such custom descriptor sets yield reliable and reproducible results.

In this paper we report Mold^2, a software package for calculating 779 molecular descriptors, solely from 2D chemical structures. Descriptors programmed in Mold^2 are technical implementations of well-known molecular descriptors reported in literature, none are novel in concept. Therefore, they may partially overlap with ones in some commercial software packages. Mold^2 molecular descriptors are easily and quickly calculated with no missing values, a common problem with most existing commercial systems. Comparisons with molecular descriptors calculated using several other commercial software packages demonstrated that Mold^2 molecular descriptors yield models that are comparable in quality when used with the classification software Decision Forest[48–50] for several reported data sets.[44–47] We show through Shannon entropy analysis that the Mold2 descriptors convey a similar amount of information as the other tested descriptor sets. Mold^2 is written in C++ on Windows XP system and is freely available to the public.

MOLD^2 DESCRIPTORS

A number of molecular descriptors have been reported in the literature and can be grouped into three categories:[11,16,35] 1D, 2D, and 3D. 1D molecular descriptors present bulk properties of compounds, such as the number of specific atoms, molecular weight, etc., and can be calculated solely based on a molecular formula. 2D molecular descriptors present structural information that can be computed from 2D structure of a molecule, such as the number of benzene rings, the number of hydrogen bond donors, etc. 3D molecular descriptors present structural information that has to be derived from 3D representation of a molecule, such as solvent accessible surface area with positive partial charge in the structure.

It has been argued that 3D molecular descriptors generally perform better than 2D molecular descriptors in QSAR and other applications where the 3D structure of a compound, including absolute stereochemistry, is critical for binding to a receptor. However, the opposite observations have been reported in comparative studies using different methods,[51–53] and in the comparative analysis presented here (see section titled "Evaluation of Descriptors"), indicating that 2D descriptors could perform equivalently to 3D descriptors in most applications. An optimal set of molecular descriptors in the absolute sense is currently not available and may not exist. For many applications in QSAR and predictive toxicology, the use of simpler 2D molecular descriptors appears to be sufficient, and beneficial, given the difficulties and uncertainties associated with using biologically active conformations for 3D molecular descriptors.

The current version of Mold^2 calculates 779 1D and 2D molecular descriptors. The 779 descriptors can be grouped into 20 categories given in Table 1 based on theoretical consideration.

The 1D descriptors (sometimes called 0D descriptors in the literature) are calculated solely based on the molecular formula. The atom counts include numbers of different atoms and the total number of atoms in the molecule. Two physicochemical properties are molecular weight and average molecular weight.

The 2D descriptors are calculated from the 2D structure of a molecule, though some of them such as logP and fragment counts are called 1D descriptors in the literature.

The 2D descriptor's counts of atoms are different from those in the 1D descriptors since different types of atoms are counted. That is, all types of carbon atoms are considered as the same in 1D descriptors because a 1D molecular formula does not distinguish among them. Types of carbon atoms in the 2D descriptors are distinguished based on hybridization status, such as primary carbon, tertiary carbon on ring structure, unsubstituted aromatic carbon, and so on. Mold^2 includes most atom types that are present in organic compounds.

The second type of 2D descriptors is related to the bond information such as numbers of single bonds, double bonds, aromatic bonds, rotatable bonds, and so on.

There are 106 functional groups available from Mold^2: examples are a carboxylic group on an aromatic ring, a tertiary amide, an aromatic aldehyde, an aliphatic hydroxylamine, a sulfonic acid, and so forth.

Table 1. Molecular Descriptors in Mold²

class	subclass	number of descriptors	example of descriptors
1D	counts for atoms	105	number of O atoms
	chemical physical property	2	molecular weight
2D	counts for atoms	80	number of ring tertiary C
	counts for bonds	9	number of rotatable bonds
	counts for functional groups	106	number of carboxylic (aromatic)
	chemical physical property	16	logP
	structural features	13	number of 5 member rings
	2D autocorrelation	96	Moran coefficient
	Balaban index	12	normalized centric index
	connectivity index	36	Randic connectivity index
	detour index	24	cyclicity index
	distance (topological) index	73	average atom eccentricity
	Eigen value based descriptors	88	folding degree index
	information content	45	mean information content
	Kier index	14	Kier flexibility
	molecular walk counts	13	total walk count
	Schultz index	4	reciprocal Schultz index
	topological charge index	21	mean topological charge
	Wiener index	17	normalized Wiener index
	Zagreb index	5	quadratic index

The physicochemical properties are calculated from the 2D structure of a compound, which are not similar to the ones in 1D descriptors that are obtained only based on the molecular formula. The octanol/water partition coefficient parameter, logP,[9] van der Waals volume, and the sum of Pauling atomic polarizabilities are examples of this subclass of 2D descriptors.

Structural features are specific structural components whose chemical and biological functionality is not explicitly understood, such as the number of independent rings, circuits, aromatic rings, three-membered rings, and so on.

The 2D autocorrelation descriptors[33,36] $A(d)$, are calculated from the 2D structure based on the autocorrelation function

$$A(d) = \sum_{j=1}^{a}\sum_{i=1}^{a} \sigma(d_{ij} - d)p_i p_j$$

$$\sigma = \begin{cases} 0 (d_{ij} \neq d) \\ 1 (d_{ij} = d) \end{cases} \quad (1)$$

where d is a topological distance and can be any number between 1 and the maximum of distance in a molecule, σ is a function of the variable d_{ij} (the topological distance between atoms i and j), a is the number of atoms in the molecule, and p_i and p_j are the properties of atoms i and j. ATS (autocorrelation of topological structures) and the Moran coefficient (general spatial autocorrelation index) are the members of this subclass of descriptors.

The Balaban index descriptors[8,26] are obtained from the Balaban distance connectivity index J that is calculated using the following formula

$$J = \frac{B}{C+1}\sum_{k=1}^{B}\sqrt{(v_i v_j)_k} \quad (2)$$

where v_i and v_j are the vertex distance degrees of two atoms connected by bond k, B is the number of bonds of the molecule, and C is the cyclomatic number.

The connectivity index descriptors[5,7,10,30,39] are calculated from the topological structure of a molecule by using a formula similar to the index J. The total connectivity index, local connectivity index, and Randic connectivity index are the examples of such connectivity index descriptors.

Descriptors such as the cyclicity index and average cyclicity index are part of the detour index[10,13] that is calculated from the detour distance matrix of a molecule. The detour matrix (or maximum path matrix) $[\Delta]_{ij}$ is a square symmetric matrix

$$[\Delta]_{ij} = \begin{cases} \Delta_{ij} = {}^{\max}p_{ij} & (i \neq j) \\ 0 & (i = j) \end{cases} \quad (3)$$

where its element of row i and column j ($^{\max}p_{ij}$) is the maximum number of edges between nodes i and j and is zero from a node to itself.

The topological distance index[5,8,33,39] descriptors are derived from the distance matrix **D** (or vertex distance matrix) of a molecule.

$$[D]_{ij} = \begin{cases} d_{ij} = {}^{\min}p_{ij} & (i \neq j) \\ 0 & (i = j) \end{cases} \quad (4)$$

The distance matrix is a square symmetric matrix summing up the topological distance information between all pairs of atoms. Its elements represent the shortest paths in the number of edges from an atom to another, and by convention it is zero from an atom to itself. The average atom connectivity index and the Rouvray index (total connectivity index) are examples of this type of descriptors.

Descriptors extracted from the eigenvalues[21] of an adjacent matrix and other matrices of a molecule are classified into the group of eigenvalue based descriptors, including the Lovasz-Pelikan index, the folding degree index, the characteristic root index, and so on.

The information content[56] descriptors are derived from the information content of a molecule (I_c). I_c is used to measure the degree of diversity of the atoms or bonds in a molecule and defined by the formula

$$I = \sum_{c=1}^{C} n_c \log_2 n_c \quad (5)$$

where C is the number of different types of atoms or bonds, and n_c is the number of atoms or bonds of the cth type. Mean information content, mean information content on edge equality, and the redundancy index are some examples.

The Kier index descriptors[6,12,31] are derived from the Kier shape indices, a set of topological shape indices defined in terms of the number of graph vertices and the number of paths with fixed length m (m=1, 2, 3) in the H-deleted molecular graph. The descriptors include Kier shape descriptors, Kier steric descriptors, Kier flexibility descriptors, and so on.

Descriptors related to molecular walk counts[7,14] are calculated based on the graph walks. They are extracted from

the adjacency matrix of a molecule. Total walk count, weighted walk degrees, and walk connectivity indices are examples.

The subclass of Schultz index *SI* descriptors[18] are calculated based on the formula

$$SI = \sum_{i=1}^{a} [(M + D) \bullet v]_i \quad (6)$$

where a is the number of atoms or bonds in a molecule, M is the adjacency matrix (nodes or edges), D is the distance matrix (nodes or edges), and v is the vertex or edge degree vector.

Descriptors related to the topological charge index[15,23] are derived from the adjacency matrix and distance matrix of a molecule, which estimate the charge transfer between pairs of atoms, and therefore the global charge transfer in the molecule, including the total topological charge, the mean topological charge, and so on.

The Wiener index[19] was originally defined only for acyclic graphs. The Wiener index descriptors in Mold^2 are derived from the Wiener matrix modified from the distance matrix of a molecule. Examples include the normalized Wiener index, the Wiener index degree, and the resistance distance hyper-Wiener index.

The Zagreb index descriptors[33] are derived from the vertex degree of atoms of a molecule, including the quadratic index and the binormalized quadratic index.

GENERATION OF MOLD^2 DESCRIPTORS

Molecular descriptors are generated from 2D molecular structures of chemicals. Mold^2 accepts an SDfile of the molecules for which descriptors are to be calculated. Other formats for representing molecular structure have to be converted to an SDfile format for use by Mold^2. The chemical structures in an SDfile are processed sequentially. Prior to descriptor calculation, Mold^2 preprocesses the input structure to provide a check whether, in fact, the structure is as expected and correct. The preprocessing is a tedious process of standardizing the structures, which includes counterion removal, adding hydrogen atoms to heavy atoms, correcting errors of an input structure, and transforming the initial Molfile to the connection table that is operated on by Mold^2. The main part of Mold^2 is a module for computing the 779 descriptors. A set of generalized functions is used to speed up calculations. Among the most important ones is a module for the perception of the smallest set of smallest rings (SSSR) that is adopted from a previously reported algorithm.[54] The related perception of the aromatic ring system was implemented based on the same algorithm. The functions important for generating walk matrix, distance matrix, and path matrix were developed by modifying the publicly available algorithm[54] while adopting the canonical representation of a structure.[55] Descriptors are generated using a set of subroutines, where related descriptors are simultaneously calculated in the same subroutine to minimize redundant calculations. Once all 779 molecular descriptors for a molecule are calculated, results are output to a file (not kept in memory); molecules are serially processed until all molecules in the SDfile are processed. The processing status and errors are recorded in a log file.

Mold^2 was developed in C++ on the Windows XP system. It is operated by the command line. The descriptors of all molecules in a single SDfile (exported from any database) can be calculated in a single run. The API (Application Program Interface) for Mold^2 is in development. The executable file for Mold^2 is publicly available.

EVALUATION OF MOLD^2 DESCRIPTORS

Selection of proper molecular descriptors is an essential step for QSAR/QSPR modeling. Accordingly, evaluation of the performance of molecular descriptors is necessary for guiding their proper use.

There are three ways of evaluating the utility of descriptors. The first is to assess information that is presented in a data set represented by a set of descriptors. Here the variance of a descriptor among molecules in a data set is a measure of the information represented by the descriptor. In general, higher variance in the descriptors corresponds to a high probability of developing a valid model using the descriptors. The second is to make sure that there are not any redundant descriptors and not many highly correlated descriptors. The third is with a comparative analysis of sets of descriptors when the same modeling approach is applied to the same data set; the differences in modeling results can then be used to differentiate the efficacy of two sets of descriptors. The Mold^2 descriptors were evaluated using all of the three approaches. Specifically, Mold^2 descriptors were compared with descriptors from three commercial software packages using information entropy analysis, analysis of correlations between descriptors, and Decision Forest classification on several reported data sets.

Information Content by Shannon Entropy Analysis. The concept of Shannon entropy,[56] also called information entropy, has played a central role in information theory. It can be used as a measure of uncertainty. In the approach, the entropy of a random variable is associated with its probability distribution, which is formulated as

$$H_n(p_1, p_2 \cdots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (7)$$

where p_i is the probability of outcome i. In information theory, the entropy is conceptually the actual amount of (information theoretic) information in a data set. It is reasonable to use the entropy of descriptors as an estimation or measure of the structural information contained in the descriptors of the data set. Golden et al.[57] used Shannon entropy to analyze different descriptors by ranking them according to variance, with high variance descriptors preferred for discriminating compounds.

Shannon entropy analysis was conducted for comparing Mold^2 descriptors with three sets of descriptors generated from the commercial software packages (i.e., Molconn-Z, Dragon, and Cerius^2) by using four published data sets, listed in Table 2 as NCTR_ER,[44] NCTR_AR,[45] ER_comb,[46] and EPA.[47] First, descriptors not having a value across all the chemicals in a data set were discarded. For example, the descriptor for the number of P atoms was discarded if any compound in a data set had no P atom. Next, descriptors were binned for each descriptor in the data sets and the probability distributions calculated using 20 even bins that spread from the minimum to the maximum values of the

Table 2. Shannon Entropy Analysis Results

data set		NCTR_AR	NCTR_ER	ER_Comb	EPA
size (compounds)		202	232	1086	57453
Cerius^2	descriptors	205	197	228	[a]
	mean entropy	1.962	1.995	1.5983	
Dragon	descriptors	616	604	671	[a]
	mean entropy	2.453	2.431	1.996	
Molconn_Z	descriptors	331	312	378	450
	entropy	1.792	1.798	1.021	0.801
Mold^2	descriptors	590	578	626	742
	entropy	2.408	2.380	2.058	1.34

[a] The licenses of Cerius^2 and Dragon are not available in our laboratories now, thus the descriptors for EPA data set were not analyzed.

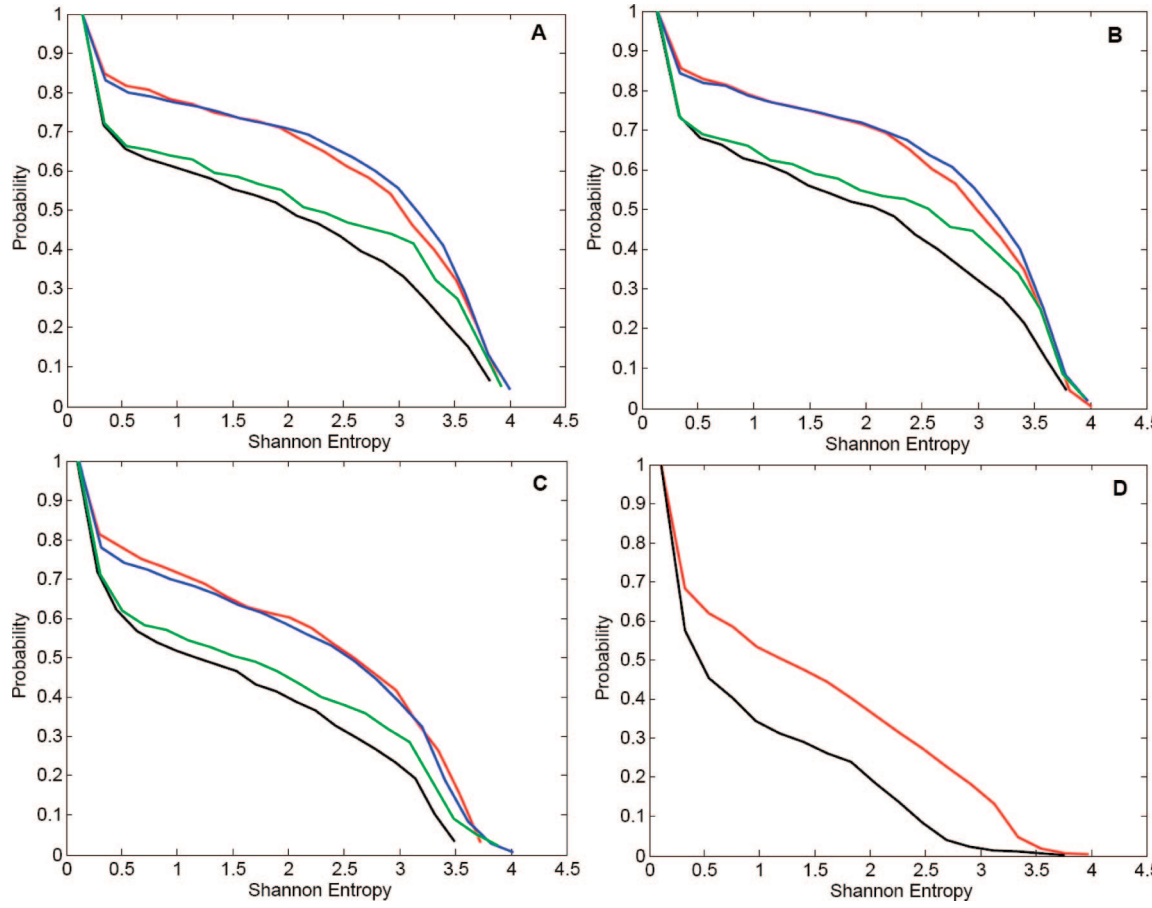

Figure 1. Reverse cumulative probability versus the Shannon entropy of descriptors for data sets NCTR_AR (**A**), NCTR_ER (**B**), ER_Comb (**C**), and EPA (**D**). The *x*-axis represents the Shannon entropies of descriptors, while the *y*-axis is the reverse cumulative probability of descriptors having Shannon entropy greater than or equal to the *x* value. The curves of Mold^2 descriptors are colored in red; Cerius^2 in green; Dragon in blue; and Molconn-Z in black.

descriptor. The binning approach allows different units and values to be comparable, a necessity since the number and type of descriptors from different software packages vary substantially. Then, the comparison of descriptor sets was done assuming that mean Shannon entropy was proportional to the average information encoded in different sets of descriptors for the same data set. As shown in Table 2, for all four data sets, the mean Shannon entropy of descriptors from Mold^2 is comparable with descriptors from Dragon and slightly higher than those from Cerius^2 and Molconn-Z. This provides confidence that the Mold^2 descriptors are equal or more informative than the ones from the compared commercial software packages.

The mean Shannon entropy only estimates the average information presented in the descriptors. The distribution of Shannon entropy is another important property of the descriptors, as it is expected that the higher entropy descriptors are, the better a data set is represented by the descriptors. Thus, the distributions of Shannon entropy for all sets of descriptors on the four data sets were plotted as the reverse cumulative probability distributions against Shannon entropy. The results are shown in Figure 1. The curves in the figure can be read as follows: a point (*x*, *y*) in a curve indicates that the probability (percentage) of descriptors (*y*) has Shannon entropy greater than or equal to *x*. For example, for the NCTR_ER data set (Figure 1**B**), the probability of Mold^2 descriptors with Shannon entropy greater than or equal to 2 is 0.721, 0.551 for Cerius^2, 0.723 for Dragon, and 0.505 for Molconn-Z. The slightly larger probability value in the high entropy region for Mold^2 across all four data sets

Table 3. Shannon Entropy of Top 20 Descriptors

NCTR_AR				NCTR_ER				ER_Comb				EPA	
Mold^2	Molcolm-Z	Cerius^2	Dragon	Mold^2	Molcolm-Z	Cerius^2	Dragon	Mold^2	Molcolm-Z	Cerius^2	Dragon	Mold^2	Molcolm-Z
4.02	3.92	4.03	4.11	4.12	3.89	4.05	4.08	3.82	3.58	3.98	4.13	4.30	4.30
4.02	3.90	3.98	4.05	3.93	3.87	4.03	4.07	3.82	3.58	3.92	3.99	4.30	3.79
3.97	3.90	3.93	4.03	3.89	3.84	3.91	3.99	3.81	3.56	3.89	3.97	4.17	3.79
3.94	3.90	3.92	4.02	3.86	3.79	3.91	3.98	3.80	3.53	3.86	3.95	3.68	3.41
3.93	3.86	3.91	4.02	3.85	3.79	3.87	3.97	3.79	3.47	3.78	3.91	3.65	3.22
3.93	3.85	3.90	4.00	3.82	3.79	3.85	3.97	3.78	3.46	3.78	3.91	3.60	3.09
3.93	3.85	3.87	3.99	3.77	3.78	3.84	3.92	3.78	3.44	3.71	3.90	3.58	2.95
3.92	3.83	3.86	3.99	3.75	3.77	3.83	3.92	3.76	3.41	3.71	3.89	3.57	2.95
3.91	3.82	3.86	3.98	3.75	3.76	3.81	3.89	3.72	3.41	3.71	3.86	3.54	2.92
3.90	3.80	3.84	3.97	3.75	3.75	3.75	3.89	3.72	3.41	3.70	3.83	3.53	2.81
3.90	3.80	3.83	3.97	3.74	3.75	3.74	3.89	3.70	3.41	3.65	3.81	3.48	2.80
3.90	3.78	3.81	3.96	3.74	3.70	3.74	3.85	3.70	3.41	3.63	3.80	3.47	2.78
3.89	3.77	3.81	3.95	3.74	3.70	3.73	3.85	3.69	3.40	3.55	3.78	3.47	2.76
3.88	3.77	3.80	3.94	3.74	3.70	3.70	3.83	3.68	3.38	3.51	3.77	3.45	2.69
3.87	3.77	3.80	3.93	3.73	3.67	3.70	3.83	3.67	3.38	3.44	3.75	3.42	2.69
3.87	3.77	3.79	3.93	3.73	3.67	3.67	3.82	3.67	3.38	3.41	3.74	3.42	2.66
3.87	3.77	3.78	3.93	3.72	3.65	3.67	3.82	3.66	3.37	3.41	3.74	3.41	2.58
3.86	3.75	3.78	3.93	3.72	3.65	3.64	3.81	3.65	3.36	3.41	3.72	3.40	2.58
3.86	3.73	3.77	3.93	3.72	3.63	3.64	3.79	3.63	3.36	3.41	3.71	3.40	2.57
3.86	3.73	3.75	3.93	3.72	3.61	3.64	3.79	3.62	3.36	3.40	3.71	3.38	2.57

indicates Mold^2 descriptors encode sufficient information and are no worse than those from the compared commercial software. The entropies of top 20 descriptors listed in Table 3 indicate that the most informative descriptors in the compared sets are equivalent.

Correlations between Descriptors. It is not preferable that there are many redundant or highly correlated descriptors in a set of descriptors. All the compared descriptors sets have no completely redundant descriptors. Correlation coefficients (r) between the descriptors in each of the compared descrip-

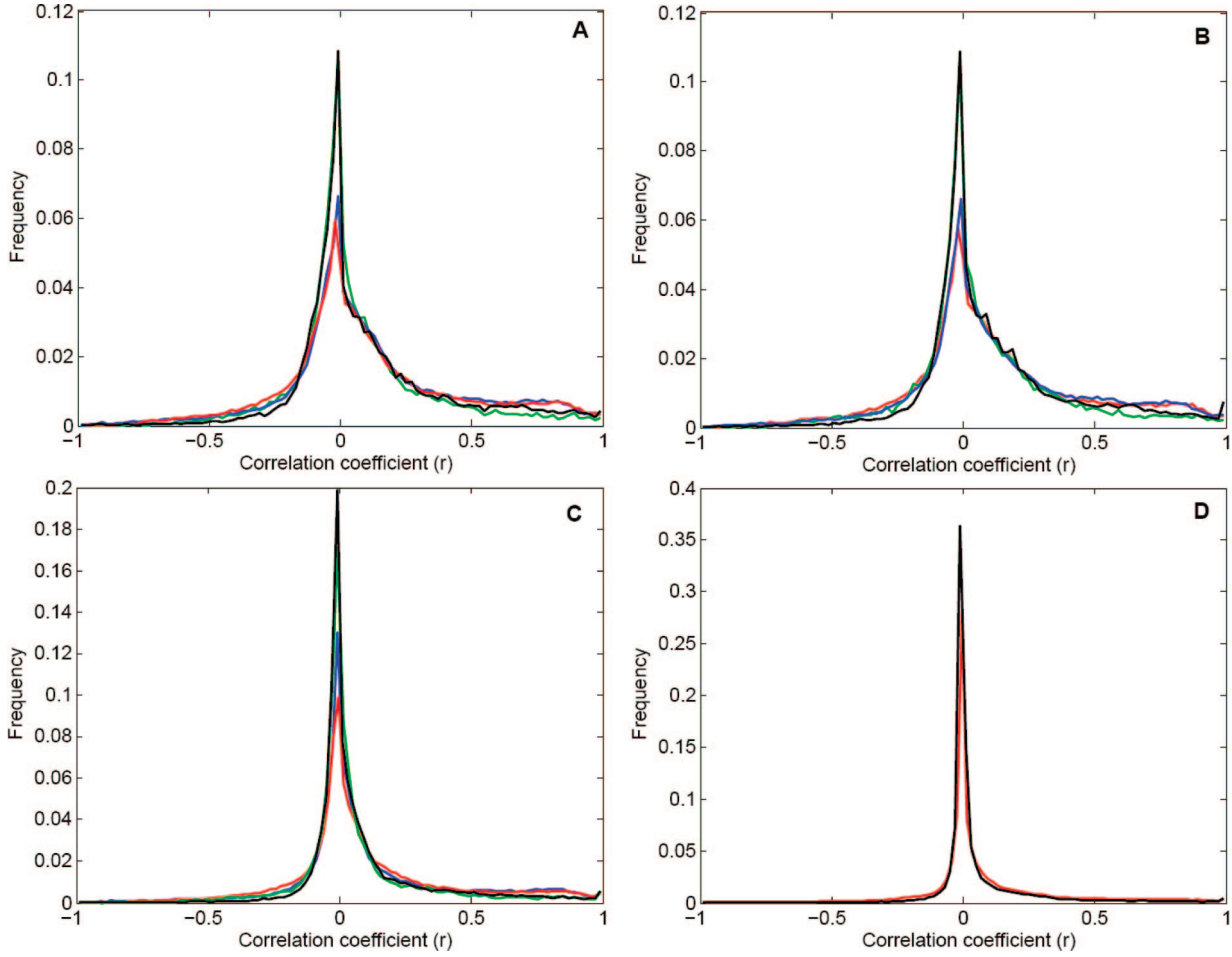

Figure 2. Distribution of correlation coefficients (r) among descriptors for data sets NCTR_AR (**A**), NCTR_ER (**B**), ER_Comb (**C**), and EPA (**D**). The *x*-axis represents the correlation coefficients (r) among descriptors, while the *y*-axis is the frequency of r. The curves of Mold^2 descriptors are colored in red; Cerius^2 in green; Dragon in blue; and Molconn-Z in black.

Table 4. Prediction Results of 100 Runs of 10-Fold Cross-Validations (%)

statistics	data set	Cerius^2	Dragon	Molconn_Z	Mold^2
accuracy	NCTR_AR	79.09(±1.53)	77.69(±1.73)	78.18(±1.68)	80.11(±1.75)
	NCTR_ER	80.69(±0.65)	80.98(±0.72)	80.27(±0.66)	80.90(±0.68)
	ER_Comb	80.49(±1.72)	80.20(±1.80)	81.55(±1.51)	81.69(±1.64)
sensitivity	NCTR_AR	91.01(±1.59)	88.83(±1.62)	90.01(±1.65)	90.66(±1.66)
	NCTR_ER	59.19(±1.36)	60.42(±1.57)	60.46(±1.36)	61.26(±1.67)
	ER_Comb	85.72(±2.28)	85.19(±1.91)	87.79(±1.70)	86.95(±1.75)
specificity	NCTR_AR	48.02(±4.48)	48.66(±4.35)	47.36(±4.14)	52.63(±4.54)
	NCTR_ER	90.83(±0.79)	90.75(±0.73)	89.70(±0.72)	90.28(±0.73)
	ER_Comb	73.71(±2.86)	73.73(±3.12)	73.45(±2.59)	74.88(±2.87)

The data are the average values of 100 runs. The numbers in the parentheses are the standard deviations.

tors sets were calculated for the four data sets. The frequencies of correlation coefficients (r) are plotted in Figure 2. For all of the four data sets, uncorrelated descriptors are the majority for the compared descriptors sets. The weak correlations ($0.25 < r^2 < 1$) are less than 15% for all the compared descriptors sets in all the data sets. The high correlations ($r^2 > 0.8$) are less than 1%. Furthermore, the high correlations are from descriptors with a lot of missing values which are replaced with zeros. The correlation analysis demonstrated that Mold^2 descriptors hold up to current standards as the compared sets of descriptors.

Classification Performance Using Decision Forest. Shannon entropy offers a way of assessing information content and distribution encoded in descriptors. Consequently, the comparison of different sets of descriptors using Shannon entropy only reveals comparative differences in variance of the structural information encoded in the descriptors. The information (structural variance) does not necessarily correlate with the biological activities or physicochemical properties. In other words, a better set of descriptors not only carries sufficient information but also should be biologically and physicochemically relevant. In a sense, the quality of a QSAR/QSPR model is solely dependent on the correlation of chemical structure variables with biologically or physicochemically related variables, which can be investigated through comparing classification models developed from different information-bearing descriptors.

Therefore, to compare different sets of descriptors in terms of ability to correlate structural independent variables to dependent variables, Decision Forest,[48–50] a classification method developed in our laboratories, was applied to all data sets except the EPA data set (no toxicological end points available for this data set). The descriptors to be compared were calculated from Cerius^2, Dragon, Molconn-Z, and Mold^2. Ten-fold cross-validation was used, where the data set was first randomly divided into ten equal portions, and each portion was then successively excluded from the training set and predicted by the model developed from the remaining nine portions. The prediction accuracy of the cross-validation was taken as the average of prediction accuracy results of the 10 models. Each random division of the data set into 10 portions leads to 10 specific pairs of training and test sets that could be biased in terms of prediction accuracy. Therefore, the 10-fold cross-validation was repeated 100 times to achieve a statistically unbiased estimation of predictive accuracy, sensitivity, and specificity. The performances (average prediction accuracy, sensitivity, and specificity of the 100 runs of 10-fold cross-validations as well as the corresponding standard deviations) of Decision Forest models from different data sets and descriptors are listed in Table 4. Mold^2 descriptors were found to yield slightly more accurate predictive models (at the same level of quality) than those using the descriptors from the compared commercial software. The evaluation added an additional confidence that Mold^2 descriptors, besides providing a reasonable representation of chemical structures, also correlate well with biologically or physicochemically related information for the purpose of applications to various predictive toxicology and QSAR/QSPR models.

CONCLUSION

Molecular descriptors are used to represent structures of chemicals and have played an important rule in the fields of chemoinformatics and toxicoinformatics. Diverse applications include virtual library generation and screening, similarity and diversity analysis, QSAR/QSPR, and predictive toxicology. Descriptors can be either computationally or empirically derived. The 779 Mold^2 descriptors presented here are calculated from both 1D and 2D chemical structures. They provide the scientific community with a zero-cost option for efficiently obtaining a large set of informative descriptors with wide applicability to many chemoinformatics and toxicoinformatics problems.

ACKNOWLEDGMENT

We are grateful to the reviewers for their comments and suggestions for revising and improving the paper. We thank Dr. Daniel M. Sheehan for fruitful discussions and revising the manuscript. We also thank Dr. James Fuscoe, Dr. Tao Chen, and Dr. Lei Guo for reading through the paper and their comments. The views presented in this article do not necessarily reflect those of the U.S. Food and Drug Administration.

REFERENCES AND NOTES

(1) Hammett, L. P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
(2) Hansch, C.; Maloney, P. P.; Fujita, T.; and Muir, R. M. Correlation of biological activity of phenoxyacetic a cids with Hammett substituent constants and partition coefficients. *Nature* **1962**, *194*, 178–180.
(3) Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175–5180.
(4) Taft, R. W. Polar steric substituent constants for aliphatic and o-benzoate groups from rates of esterification and hydrolysis of esters. *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128.
(5) Randic, M. On the recognition of identical graphs representing molecular topology. *J. Chem. Phys.* **1974**, *60*, 3920–3928.
(6) Hall, L. H.; Kier, L. B. Structure-activity studies using valence molecular connectivity. *J. Pharm. Sci.* **1977**, *66*, 642–644.

(7) Randic, M.; Wilkins, C. L. Graph theoretical ordering of structures as a basis for systematic search for regularities in molecular data. *J. Chem. Phys.* **1979**, *83*, 1525–1540.
(8) Balaban, A. T. Topological indices based on topological distances in molecular graphs. *Pure Appl. Chem.* **1983**, *55*, 199–206.
(9) Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83–92.
(10) Randic, M. Generalized molecular descriptors. *J. Math. Chem.* **1991**, *7*, 155–168.
(11) Arteca, G. A. *Molecular shape descriptors; In Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, U.S.A., 1991; Vol. 2, pp 191−253.
(12) Hall, L. H.; Kier, L. B. *The molecular connectivity Chi indexes and Kappa shape indexes in structure-property modeling; In Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, U.S.A., 1991; Vol. 2, pp 367−422.
(13) Buckley, F.; Harary, F. *Distance Matrix in Graphs*; Addison-Wesley: Redwood City, CA, 1990.
(14) Rücker, G.; Rücker, C. Counts of all walks as atomic and molecular descriptors. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 683–695.
(15) Gasteiger, J.; Li, X.; Rudolph, C. J.; Sadowski, J.; Zupan, J. Representation of molecular electrostatic potentials by topological feature maps. *J. Am. Chem. Soc.* **1994**, *116*, 4608–4620.
(16) Todeschini, R.; Lasagni, M.; Marengo, E. New molecular descriptors for 2D- and 3D-structures. Theory. *J. Chemom.* **1994**, *8*, 263–273.
(17) Rücker, C.; Rücker, G. Mathematical relation between extended connectivity and eigenvector coefficients. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 534–538.
(18) Schultz, H. P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.
(19) Diudea, M. V. Wiener and hyper-Wiener numbers in a single matrix. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 833–836.
(20) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. Quantum-chemical descriptors in QSAR/QSPR studies. *Chem. Rev.* **1996**, *96*, 1027–1043.
(21) Burden, F. R. A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. *Quant. Struct.-Act. Relat.* **1997**, *16*, 309–314.
(22) Ferguson, A. M.; Heritage, T. W.; Jonathon, P.; Pack, S. E.; Phillips, L.; Rogan, J.; Snaith, P. J. EVA: A new theoretically based molecular descriptor for use in QSAR/QSPR analysis. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 143–152.
(23) Galvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 520–525.
(24) Sjöberg, P. MOLSURF-A generator of chemical descriptors for QSAR. In *Computer-Assisted Lead Finding and Optimization*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Wiley-VCH: Weinheim, Germany, 1997; pp 81−92.
(25) Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113–119.
(26) Balaban, A. T. Topological and stereochemical molecular descriptors for databases useful in QSAR, similarity/dissimilarity and drug design. *SAR QSAR Environ. Res.* **1998**, *8*, 1–21.
(27) Brown, R. D.; Martin, Y. C. An evaluation of structural descriptors and clustering methods for use in diversity selection. *SAR QSAR Environ. Res.* **1998**, *8*, 23–39.
(28) Stanton, D. T. Evaluation and use of BCUT descriptors in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11–20.
(29) Dearden, J. C.; Ghafourian, T. Hydrogen bonding parameters for QSAR: Comparison of indicator variables, hydrogen bond counts, molecular orbital and other parameters. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 231–235.
(30) Randic, M.; Basak, S. C. Optimal molecular descriptors based on weighted path numbers. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 261–266.
(31) Kier, L. B.; Hall, L. H. *Molecular Structure Description. The Electrotopological State*; Academic Press: London, U.K., 1999.
(32) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprint. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569–574.
(33) Ivanciuc, O.; Ivanciuc, T. *Matrices and structural descriptors computed from molecular graphs distances; In Topological Indices and Related Descriptors in QSAR and QSPR*; Devillers, J., Balaban, A. T., Eds.; Gordon & Breach: Amsterdam, The Netherlands, 2000; pp 221−277.
(34) Ivanciuc, O.; Taraviras, S. L.; Cabrol-Bass, D. Quasi-orthogonal basis sets of molecular graph descriptors as a chemically diversity measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126–134.
(35) Xue, L.; Bajorath, J. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. *Comb. Chem. High Throughput Screening* **2000**, *3*, 363–372.
(36) Krenkel, G.; Castro, E. A.; Toropov, A. A. Improved molecular descriptors based on the optimization of correlation weights of local graph invariants. *Int. J. Mol. Sci.* **2001**, *2*, 57–65.
(37) Padron, J. A.; Carrasco, R.; Pellon, R. F. Molecular descriptor based on a molar refractivity partition using Radic-type graph-theoretical invariant. *J. Pharm. Pharmaceut. Sci.* **2002**, *5*, 258–265.
(38) Faulon, J. L.; Visco, D. P.; Pophale, R. S. The signature molecular descriptor. 1. Using extended valence sequences in QSAR and QSPR studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707–720.
(39) Randic, M.; Basak, N.; Plavsic, D. Novel graphical index and distance-based molecular descriptors. *Croat. Chem. Acta* **2004**, *77*, 251–257.
(40) Jurs, P. C.; Chou, J. T.; Yuan, M. Computer-assisted structure-activity studies of chemical carcinogens. A heterogeneous data set. *J. Med. Chem.* **1979**, *11*, 179–186.
(41) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. *Chem. Soc. Rev.* **1995**, *24*, 279–287.
(42) Mekenyan, O.; Karabunarliev, S.; Bonchev, D. The microcomputer OASIS system for predicting the biological activity of chemical compounds. *Comp. Chem.* **1990**, *14*, 193–200.
(43) Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Disc. Appl. Math.* **1988**, *19*, 17–44.
(44) Blair, R.; Fang, H.; Branham, W. S.; Hass, B.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi., L.; Perkins, R.; Sheehan, D. M. Estrogen receptor relative binding affinities of 188 natural and xenocompounds: structure diversity of ligands. *Toxicol. Sci.* **2000**, *54*, 138–153.
(45) Fang, H.; Tong, W.; Branham, W.; Moland, C. L.; Dial, S. L.; Hong, H.; Xie, Q.; Perkins, R.; Owens, W.; Sheehan, D. M. Study of 202 Natural, Synthetic and Environmental Chemicals for Binding to the Androgen Receptor. *Chem. Res. Toxicol.* **2003**, *16*, 1338–1358.
(46) Tong, W.; Xie, Q.; Hong, H.; Fang, H.; Shi, L.; Perkins, R. Assessment of Prediction Confidence and Domain Extrapolation of Two Structure-activity Relationship Models for Predicting Estrogen Receptor Binding Activity. *EHP Toxicogenomics* **2004**, *112*, 1249–1254.
(47) Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J. D.; Branham, W.; Sheehan, D. M. Prediction of estrogen receptor binding for 58,000 compounds using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* **2002**, *110*, 29–36.
(48) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525–531.
(49) Hong, H.; Tong, W.; Perkins, R.; Fang, H.; Xie, Q.; Shi, L. Multiclass decision forest-a novel pattern recognition method for multiclass classification in microarray data analysis. *DNA Cell Biol.* **2004**, *23*, 685–694.
(50) Hong, H.; Tong, W.; Xie, Q.; Fang, H.; Perkins, R. An *in silico* ensemble method for lead discovery: decision forest. *SAR QSAR Environ. Res.* **2005**, *16*, 339–347.
(51) McGregor, M. J.; Pallai, P. V. Clustering large databases of compounds using the MDL 'keys' as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
(52) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.
(53) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1211–1225.
(54) Hong, H.; Xin, X. ESSESA, an expert system for structure elucidation from spectral analysis Part II. Novel algorithm of perception of the linear independent smallest set of smallest rings. *Anal. Chim. Acta* **1992**, *262*, 179–191.
(55) Hong, H.; Xin, X. ESSESA, an expert system for structure elucidation from spectra. 4. Canonical representation of structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 730–734.
(56) Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423.
(57) 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.

CI800038F