Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets

Roberto Todeschini,*[†] Viviana Consonni,[†] Hua Xiang,[‡] John Holliday,[‡] Massimo Buscema,[§,⊥] and Peter Willett[‡]

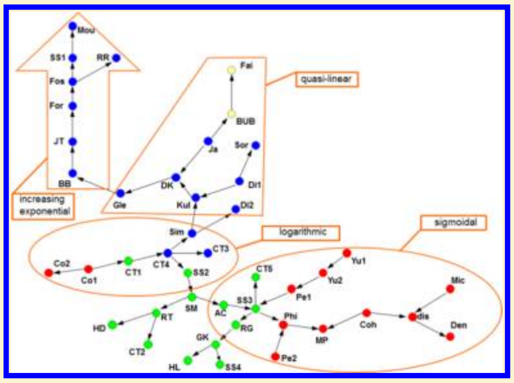
[†]Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza 1, 20126, Milano, Italy

[‡]Information School, University of Sheffield, Sheffield S1 4DP, United Kingdom

[§]SEMEION, via Sersale 117,00128, Roma, Italy

[⊥]Department of Mathematical and Statistical Sciences, University of Colorado, 1250 14th Street, 80217-3364 Denver, CO, USA

ABSTRACT: This paper reports an analysis and comparison of the use of 51 different similarity coefficients for computing the similarities between binary fingerprints for both simulated and real chemical data sets. Five pairs and a triplet of coefficients were found to yield identical similarity values, leading to the elimination of seven of the coefficients. The remaining 44 coefficients were then compared in two ways: by their theoretical characteristics using simple descriptive statistics, correlation analysis, multidimensional scaling, Hasse diagrams, and the recently described atemporal target diffusion model; and by their effectiveness for similarity-based virtual screening using MDDR, WOMBAT, and MUV data. The comparisons demonstrate the general utility of the well-known Tanimoto method but also suggest other coefficients that may be worthy of further attention.



■ INTRODUCTION

The data in many application domains consist of strings in which binary variables are used to reflect either the presence or absence of certain attributes. For example, in archeology, binary data may reflect the fact that some particular type of artifact was or was not found in a specific grave; in psychology, binary data may indicate if people do or do not possess a certain psychological trait; in ecology, objects could be regions in which certain species do or do not occur; and in social network studies, individuals can be described by binary variables indicating whether or not they are involved in specific types of relationship. In chemistry, molecules are frequently described by binary strings representing the presence/absence of certain features in the 2D or 3D structure, or of signal in instrumental spectra. In the present paper, we focus on the former, extremely common type of representation, which is normally referred to as a fingerprint. Specifically, we investigate the use of 2D fingerprints for the calculation of intermolecular structural similarity, which plays a key role in many applications in chemoinformatics. For example, read-across is a recent QSAR strategy that has aroused interest for chemical risk assessment according to the REACH regulations and that seeks to detect chemicals whose physicochemical, toxicological, and ecotoxicological properties are likely to be similar or to follow a regular pattern as a result of structural similarity to molecules for which the properties of interest are already known. Fingerprints play a fundamental role in evaluating similarity patterns to determine structurally similar molecules and, hence, in performing property prediction calculations for previously untested molecules. Another application, and one that is studied in some detail below, is similarity-based virtual screening. Virtual screening involves ranking a chemical database in order of decreasing probability of activity, so that compound acquisition or synthesis and biological testing can be focused on those database structures that have high a priori probabilities of activity.[1−4] Similarity searching is one of the most common forms of virtual screening and involves the use of a bioactive reference structure that is compared with each of the structures in a database to determine the intermolecular structural similarity in each case. The database structures are then ranked in order of decreasing similarity and the top-ranked molecules, or nearest neighbors, passed on for further investigation. The basis for similarity searching is the similar property principle, which states that molecules that are structurally similar are expected to have similar properties.[5,6] The nearest neighbors in a similarity search for a bioactive reference structure are hence likely to exhibit the desired bioactivity, given an appropriate similarity measure.

A chemical similarity measure has three components: the structural representation that is used to characterize the reference structure and each of the database structures by some set of structural features; a weighting scheme that can be applied to the basic representation to reflect differing degrees of importance for each of the features comprising the representation; and the

Received: June 8, 2012

ACS Publications © XXXX American Chemical Society

similarity coefficient that is used to quantify the degree of resemblance between two (possibly weighted) representations. Many different types of structural representations have been described for similarity searching,[7−9] but in this paper we restrict ourselves to binary chemical fingerprints, which have been by far the most widely used for this purpose. We also restrict ourselves to unweighted fingerprints, where all of the substructural features encoded in the fingerprint are assumed to be of equal importance. The focus of this paper is the third component, where many different types of coefficient have been proposed in the literature to compute the similarity of two binary vectors. Such coefficients, often called association coefficients, reflect in some way the association or resemblance of two objects and are used, e.g., to compute a similarity matrix of pairwise resemblances that then forms the input to multivariate data analysis methods such as multidimensional scaling or cluster analysis.

Given that many different association coefficients are available, the question arises as to which one (or ones) are most suitable for a particular application domain, this in turn requiring an understanding of their theoretical and practical properties. The Tanimoto coefficient has found to be effective in many applications in chemoinformatics, including virtual screening, but it is clearly of interest to determine whether other coefficients might yield a higher level of screening effectiveness, and there have been many previous comparisons of association coefficients for this purpose.[7,10−12] None, however, have involved either the number or the range of coefficients considered here: specifically, this paper considers a total of 51 binary similarity coefficients in a comparative study that seeks to evaluate their properties, relationships, diversity, and effectiveness in virtual screening. Two types of study are presented: first, an exploratory study of the mathematical properties of the coefficients and their basic relationships, using a simulated data set and both classical linear and novel nonlinear methods of multivariate data analysis; second, an extended comparative study of the effectiveness of these coefficients when used for similarity-based screening of real data sets for which bioactivity data were available.

■ SIMILARITY COEFFICIENTS

Theory. Let us start from the formal definition of similarity. Let X be a set. Then a function s: X × X → $\mathbb{R}$ is called a similarity on X if s is non-negative, symmetric and if $s(x,y) \leq s(x,x)$ holds for all $x,y \in$ X, with equality if and only if $x = y$.

To deal with binary descriptors, i.e., variables whose values are either one or zero (presence or absence), several similarity coefficients have been proposed in the literature,[13−17] and they can all be described as follows. Let two objects be described by the binary vectors $\mathbf{x}$ and $\mathbf{y}$, each comprised of p variables with values 0/1. The common association coefficients are calculated from the data reported in Table 1, where a, b, c, and d are the frequencies of the events ($x = 1$ and $y = 1$), ($x = 1$ and $y = 0$), ($x = 0$ and $y = 1$), and ($x = 0$ and $y = 0$), respectively, in the pair of binary vectors describing the two objects; p is the total number of variables, equal to $a + b + c + d$, which is the length of each binary vector.

Table 1. Frequency Table of the Four Possible Combinations for Two Binary Variables

	$y = 1$	$y = 0$	
$x = 1$	a	b	$a + b$
$x = 0$	c	d	$c + d$
	$a + c$	$b + d$	p

The frequency table can be read as follows: a is the number of common "presences" of the attributes and d is the number of common "absences" in $\mathbf{x}$ and $\mathbf{y}$; $a + b$ is the number of presences of the attributes in $\mathbf{x}$ and $a + c$ is the number of presences of the attributes in $\mathbf{y}$. The diagonal entries a and d hence give information about the similarity between the two vectors, while the entries b and c give information about their dissimilarity.

A simple example is presented below to show how the frequencies a, b, c, and d are calculated. Let two objects be represented by the vectors $\mathbf{x}$ and $\mathbf{y}$, each described by ten binary variables, i.e., $p = 10$:

$\mathbf{x}$: 1 1 0 1 1 0 0 0 1 1

$\mathbf{y}$: 1 0 1 0 1 0 0 0 1 1

Then, a, b, c, and d take the following values:

$$\begin{vmatrix} a = 4 & b = 2 \\ c = 1 & d = 3 \end{vmatrix}$$

with their sum, $a + b + c + d$, equal to 10, i.e., the total number of binary variables.

An alternative representation of binary similarity measures is based on set theory and is well-described by Batagelj.[13] Let $\mathbf{xy}$ be the scalar product of a set X of p binary variables $\mathbf{x}$ and $\mathbf{y}$

$$\mathbf{xy} = \sum_{i=1}^{p} x_i y_i \qquad x_i, y_i \in \mathrm{X} = [0, 1]$$

and let $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ be the complementary vectors defined as follows:

$$\bar{\mathbf{x}} = \mathbf{1} - \mathbf{x} \text{ and } \bar{\mathbf{y}} = \mathbf{1} - \mathbf{y}$$

Then, for any pair of vectors $\mathbf{x}$ and $\mathbf{y}$, the following counters can be defined as follows:

$a = \mathbf{xy}$ numbers of attributes which $\mathbf{x}$ and $\mathbf{y}$ share

$b = \mathbf{x}\bar{\mathbf{y}}$ numbers of attributes which $\mathbf{x}$ has and $\mathbf{y}$ lacks

$c = \bar{\mathbf{x}}\mathbf{y}$ numbers of attributes which $\mathbf{x}$ lacks and $\mathbf{y}$ has

$d = \bar{\mathbf{x}}\bar{\mathbf{y}}$ numbers of attributes which both $\mathbf{x}$ and $\mathbf{y}$ lack

where, again, $a + b + c + d = p$.

We define a *symmetric* similarity coefficient to be one that uses both a and d, i.e. the double-zero state (d) for two objects is treated in exactly the same way as any other pair of values and should be used when the zero state is a valid basis for comparing two objects; an *asymmetric* coefficient, conversely, ignores such double-zero attributes in the similarity calculation. Asymmetric indices are commonly used in ecology where it is well-known that

> *if a species is present at two sites, this is an indication of the similarity of these sites: but if a species is absent from two sites, it may be because the two sites are both above the optimal niche value for that species, or both are below, or else one site is above and the other is below that value. One cannot tell which of these circumstances is the correct one.*[16]
>
> Legendre and Legendre

Asymmetric indices have also been widely used in chemoinformatics, since molecules can be assumed to be similar if they have many fragments in common[18,19] but are unlikely to be regarded as highly similar when two molecules both lack large numbers of fragments. It should be noted that this partition of coefficents into families of different symmetry is the definition commonly used in the field of binary similarity coefficients,[13] but, from a closer mathematical point of view, the condition for symmetric functions is not fulfilled for a similarity if $s(\mathbf{x},\mathbf{y}) \neq s(\mathbf{y},\mathbf{x})$.

Binary Similarity Coefficients. In this paper, 51 similarity coefficients for binary variables were retrieved from the literature in order to perform an extended comparison using both simulated and real data. These coefficients are listed in Table 2. In order to enhance their comparability, coefficients having ranges other than [0, 1] were rescaled using the following linear transform:

$$s' = \frac{s + \alpha}{\beta}$$

where s is the original similarity value, s' is the rescaled function in the range [0, 1], and α and β are numerical parameters whose values are reported in columns 4 and 5 of Table 2 (where, obviously, $\alpha = 0$ and $\beta = 1$ indicate that no transformation needed to be applied to obtain the desired range).

Some typical binary coefficients are SM, JT, and Yu1 (using the abbreviations in Table 2):

$$(1)\ \mathrm{SM} = \frac{a + d}{p} \qquad (2)\ \mathrm{JT} = \frac{a}{a + b + c}$$

$$(3)\ \mathrm{Yu1} = \frac{ad - bc}{ad + bc}$$

The first one is called *simple matching* and is the simplest symmetric coefficient; the second is the Jaccard−Tanimoto coefficient, which is the most widely used asymmetric coefficient in chemoinformatics applications; and the third is the first Yule coefficient, a simple example of a correlation-based coefficient, ranging between [−1, +1] and here rescaled to lie in the range [0, 1].

The following information is reported for each coefficient in Table 2. First, the ID number, name, and bibliographic reference, formula, and α and β parameters (vide supra). The different properties of similarity coefficients are indicated in the column class by the following symbols: S for a symmetric measure (counts a and d are considered equally); A for an asymmetric coefficient (only count a is considered); I for an intermediate coefficient (i.e., one where a and d are considered, but where d is underweighted with respect to a); and Q for correlation-based coefficients transformed to lie between zero and one. Note that the symbol b indicates that the coefficient has been excluded from further analysis because it is perfectly correlated with some other similarity coefficient (vide infra).

The column metricity in Table 2 indicates if a coefficient is metric (M) or nonmetric (N) (vide infra); and the final column provides the conditions that were assumed in order to avoid singularities and to allow calculation of the similarity values for any possible combination of a, b, c, and d frequencies (i.e., considering extreme cases such as $d = p$ or $a = b = c = 0$).

Metric Properties. Among the mathematical properties of the binary similarity coefficients, particular attention was paid to their *metricity*, i.e., whether a similarity coefficient can be transformed into a metric distance. By definition, metric distances comply with the triangle inequality and those distances that do not comply with the triangle inequality are nonmetric or semimetric. The triangle inequality holds if

$$d(x, y) \leq d(x, z) + d(z, y)$$

where x, y, and z are three distinct points.

Once a similarity coefficient has been scaled between [0, 1], the value, s, can be easily transformed into a distance measure or, more generally, into a dissimilarity measure if not all of the distance axioms are fulfilled, using one of the following transformations:

$$d = 1 - s \qquad d = \sqrt{1 - s} \qquad d = \sqrt{2(1 - s^2)}$$

$$d = \arccos(s) \qquad d = -\ln s$$

After transformation into distances, it is easy to see that several similarity coefficients are nonmetric since it is likely that two objects, A and B, have a distance value larger than the sum of their distances with another object C. It follows that these similarities cannot be used directly to project objects in a metric space, before converting them into metric distances. These transformations do not induce metric distances if the mathematical condition for symmetric functions is not fulfilled for a similarity measure, i.e. $s(\mathbf{x},\mathbf{y}) \neq s(\mathbf{y},\mathbf{x})$. For some binary similarity coefficients, this condition is not fulfilled, for example, if only the parameter a or the parameter b, but not both contemporarily, appear in their definition: indeed, in this case, the values of b and c exchange their values. This happens for the coefficients Co1 (24), Co2 (25), Di1 (31), and Di2 (32). However, in this paper, the main purpose is to evaluate metricity of all similarity coefficients and these cases can be simply considered as an approximate evaluation of their dissimilarity behavior. The information related to the metricity is only used as class information in Figure 6.

The metricity of our 51 coefficients was evaluated using the first distance transformation above, and the results of this analysis are reported in the column metricity of Table 2.

DATA SETS

In order to compare binary similarity coefficients, two kinds of data were chosen, i.e. simulated and real data sets, with a different analysis approach for each kind.

Simulated Data for Exploratory Data Analysis. A simulated data set has been generated so as to mirror the size of the real data sets described below. Here, 100 000 cases were created by randomly generating quadruples of integer numbers (a, b, c, d) under the constraint $a + b + c + d = 1024$. In order to evaluate all the theoretically allowed combinations, the 4-tuples of a, b, c, and d parameters were generated by using a uniform distribution; the uniform distribution is far from the distributions of these parameters in virtual screening, but allows to study the coefficient behavior for any kind of application.

For each case, the 51 similarity coefficients were calculated and organized into a matrix of 100 000 rows and 51 columns, then reduced to 44 since 7 similarity coefficients were perfectly correlated, as discussed under Preliminary Analysis below. Each case can be thought of as the comparison of a binary vector of length 1024 bits with a reference vector of the same length.

This simulated data set was analyzed using Pearson correlation, Spearman rank correlation, and multi-dimensional scaling (MDS). In addition, a recently proposed technique called atemporal target diffusion model (ATDM) was used, allowing the exploration of nonlinear relationships between the coefficients.

Real Data for Virtual Screening. The evaluation of a virtual screening system requires sets of molecules for which

Table 2. List of the Binary Similarity Coefficients[a]

no.	symbol	name, date, ref	formula	α	β	class	metricity	conditions
1	SM	Sokal–Michener (1958),[20] Rand (1971),[21] simple matching	$s_{\mathrm{SM}} = \dfrac{a + d}{p}$	0	1	S	M	none
2	RT	Rogers–Tanimoto (1960)[8,22]	$s_{\mathrm{RT}} = \dfrac{a + d}{p + b + c}$	0	1	S	M	none
3	JT	Jaccard (1912)[23], Tanimoto (1960)[22]	$s_{\mathrm{JT}} = \dfrac{a}{a + b + c}$	0	1	A	M	$a = 0 \rightarrow s = 0$
4	Gle	Gleason (1920),[24] Dice (1945),[25] Sorenson (1948)[26]	$s_{\mathrm{Gle}} = \dfrac{2a}{2a + b + c}$	0	1	A	N	$a = 0 \rightarrow s = 0$
5	RR	Russel–Rao (1940)[27]	$s_{\mathrm{RR}} = \dfrac{a}{p}$	0	1	A	M	none
6	For	Forbes (1907)[28]	$s_{\mathrm{For}} = \dfrac{pa}{(a + b)(a + c)}$	0	p/a	A	M	$\mathrm{den} = 0 \vee a = 0 \rightarrow s = 0$
7	Sim	Simpson (1943)[29]	$s_{\mathrm{Sim}} = \dfrac{a}{\min\{(a + b), (a + c)\}}$	0	1	A	N	$\mathrm{den} = 0 \vee a = 0 \rightarrow s = 0$
8	BB	Braun–Blanquet (1932)[30]	$s_{\mathrm{BB}} = \dfrac{a}{\max\{(a + b), (a + c)\}}$	0	1	A	M	$a = 0 \rightarrow s = 0$
9	DK	Driver–Kroeber (1932),[31] Ochiai (1957),[32] cosine	$s_{\mathrm{DK}} = \dfrac{a}{\sqrt{(a + b)(a + c)}}$	0	1	A	N	$\mathrm{den} = 0 \rightarrow s = 0$
10	BUB	Baroni-Urbani–Buser (1976)[33]	$s_{\mathrm{BU1}} = \dfrac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$	0	1	I	M	$d = p \rightarrow s = 1$
11	Kul	Kulczynski (1927)[34]	$s_{\mathrm{Kul}} = \dfrac{1}{2}\cdot\left[\dfrac{a}{a + b} + \dfrac{a}{a + c}\right]$	0	1	A	N	$a = 0 \rightarrow s = 0$
12	SS1	Sokal–Sneath (1963)[35]	$s_{\mathrm{SS1}} = \dfrac{a}{a + 2b + 2c}$	0	1	A	M	$a = 0 \rightarrow s = 0$
13	SS2	Sokal–Sneath (1963)[35]	$s_{\mathrm{SS2}} = \dfrac{2a + 2d}{p + a + d}$	0	1	S	N	none
14	Ja	Jaccard (1912)[23]	$s_{\mathrm{Ja}} = \dfrac{3a}{3a + b + c}$	0	1	A	N	$a = 0 \rightarrow s = 0$
15	Fai	Faith (1983)[36]	$s_{\mathrm{Fai}} = \dfrac{a + 0.5d}{p}$	0	1	I	M	none
16	Mou	Mountford (1962)[37]	$s_{\mathrm{Mou}} = \dfrac{2a}{ab + ac + 2bc}$	0	2	A	M	$\mathrm{den} = 0 \rightarrow s = a/p$
17	Mic	Michael (1920)[38]	$s_{\mathrm{Mic}} = \dfrac{4(ad - bc)}{(a + d)^2 + (b + c)^2}$	+1	2	Q	N	$a = p \vee d = p \rightarrow s = 1$ $b + c = 0 \rightarrow s = 1$
18	RG	Rogot–Goldberg (1966)[39]	$s_{\mathrm{RG}} = \dfrac{a}{2a + b + c} + \dfrac{d}{2d + b + c}$	0	1	S	M	$a = p \vee d = p \rightarrow s = 1$
19	HD	Hawkins–Dotson (1968)[40]	$s_{\mathrm{HD}} = \dfrac{1}{2}\cdot\left(\dfrac{a}{a + b + c} + \dfrac{d}{b + c + d}\right)$	0	1	S	M	$a = p \vee d = p \rightarrow s = 1$

Table 2. continued

no.	symbol	name, date, ref	formula	α	β	class	metricity	conditions
20	Yu1	Yule (1900, 1912)[41,42]	$s_{\mathrm{Yu1}} = \dfrac{ad - bc}{ad + bc}$	+1	2	Q	N	$a = p \vee d = p \vee bc = 0 \rightarrow s = 1$
21	Yu2	Yule (1900, 1912)[41,42]	$s_{\mathrm{Yu2}} = \dfrac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}}$	+1	2	Q	M	$a = p \vee d = p \vee bc = 0 \rightarrow s = 1$
22	Fos	Fossum in Holiday et al. (2002)[11]	$s_{\mathrm{Fos}} = \dfrac{p(a - 0.5)^2}{(a + b)(a + c)}$	0	$(p - 0.5)^2/p$	A	M	den $= 0 \rightarrow s = 0$
23	Den	Dennis in Holiday et al. (2002)[11]	$s_{\mathrm{Den}} = \dfrac{ad - bc}{\sqrt{p(a + b)(a + c)}}$	$\sqrt{p}/2$	$\sqrt{p}$	Q	M	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
24	Co1	Cole (1949)[43]	$s_{\mathrm{Co1}} = \dfrac{ad - bc}{(a + c)(c + d)}$	$p - 1$	p	Q	N	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
25	Co2	Cole (1949)[43]	$s_{\mathrm{Co2}} = \dfrac{ad - bc}{(a + b)(b + d)}$	$p - 1$	p	Q	N	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
26	dis	dispersion in Choi et al. (2012)[44]	$s_{\mathrm{dis}} = \dfrac{ad - bc}{p^2}$	1/4	1/2	Q	N	$a = p \vee d = p \rightarrow s = 1$
27	GK	Goodman–Kruskal (1954)[45]	$s_{\mathrm{GK}} = \dfrac{2\min(a, d) - b - c}{2\min(a, d) + b + c}$	+1	2	S	N	$a = p \vee d = p \rightarrow s = 1$
28	SS3	Sokal–Sneath (1963)[35]	$s_{\mathrm{SS3}} = \dfrac{1}{4}\left[\dfrac{a}{a + b} + \dfrac{a}{a + c} + \dfrac{d}{b + d} + \dfrac{d}{c + d}\right]$	0	1	S	M	$a = p \vee d = p \rightarrow s = 1$ $a = 0 \wedge d = 0 \rightarrow s = 0$
29	SS4	Sokal–Sneath (1963)[35]	$s_{\mathrm{SS4}} = \dfrac{a}{\sqrt{(a + b)(a + c)}} \dfrac{d}{\sqrt{(b + d)(c + d)}}$	0	1	S	M	$a = p \vee d = p \rightarrow s = 1$ $a = 0 \vee d = 0 \rightarrow s = 0$
30	Phi	Pearson–Heron (1913)[46]	$s_{\mathrm{Phi}} = \dfrac{ad - bc}{\sqrt{(a + b)(a + c)(c + d)(b + d)}}$	+1	2	Q	M	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$ den $= 0 \rightarrow s = 0$
31	Di1	Dice (1945),[25] Wallace (1983),[47] Post–Snijders (1993)[48]	$s_{\mathrm{Di1}} = \dfrac{a}{(a + b)}$	0	1	A	N	$a = 0 \rightarrow s = 0$
32	Di2	Dice (1945),[25] Wallace (1983),[47] Post–Snijders (1993)[48]	$s_{\mathrm{Di2}} = \dfrac{a}{(a + c)}$	0	1	A	N	$a = 0 \rightarrow s = 0$
33	Sor	Sorgenfrei (1958)[49]	$s_{\mathrm{Sor}} = \dfrac{a^2}{(a + b)(a + c)}$	0	1	A	N	$a = 0 \rightarrow s = 0$
34	Coh	Cohen (1960)[25,50]	$s_{\mathrm{Coh}} = \dfrac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}$	+1	2	Q	N	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
35	Pe1	Peirce (1884)[51]	$s_{\mathrm{Pe1}} = \dfrac{ad - bc}{(a + b)(c + d)}$	+1	2	Q	N	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$
36	Pe2	Peirce (1884)[51]	$s_{\mathrm{Pe2}} = \dfrac{ad - bc}{(a + c)(b + d)}$	+1	2	Q	N	$a = p \vee d = p \rightarrow s = 1$ $b = p \vee c = p \rightarrow s = 0$

Table 2. continued

no.	symbol	name, date, ref	formula	α	β	class	metricity	conditions
37	MP	Maxwell–Pilliner (1968)[52]	$s_{\mathrm{MP}} = \frac{2(ad - bc)}{(a + b)(c + d) + (a + c)(b + d)}$	+1	2	Q	M	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
38	HL	Harris–Lahey (1978)[53]	$s_{\mathrm{HL}} = \frac{a(2d + b + c)}{2(a + b + c)} + \frac{d(2a + b + c)}{2(b + c + d)}$	0	p	S	N	$a = p \vee d = p \rightarrow s = 1$ den $= 0 \rightarrow s = 0$
39	CT1	Consonni–Todeschini (2012)[54]	$s_{\mathrm{CT1}} = \frac{\ln(1 + a + d)}{\ln(1 + p)}$	0	1	S	M	none
40	CT2	Consonni–Todeschini (2012)[54]	$s_{\mathrm{CT2}} = \frac{\ln(1 + p) - \ln(1 + b + c)}{\ln(1 + p)}$	0	1	S	N	none
41	CT3	Consonni–Todeschini (2012)[54]	$s_{\mathrm{CT3}} = \frac{\ln(1 + a)}{\ln(1 + p)}$	0	1	A	N	none
42	CT4	Consonni–Todeschini (2012)[54]	$s_{\mathrm{CT4}} = \frac{\ln(1 + a)}{\ln(1 + a + b + c)}$	0	1	A	N	none
43	CT5	Consonni–Todeschini (2012)[54]	$s_{\mathrm{CT5}} = \frac{\ln(1 + ad) - \ln(1 + bc)}{\ln(1 + p^2/4)}$	0	1	S	M	none
44	AC	Austin–Colwell (1977)[55]	$s_{\mathrm{AC}} = \frac{2}{\pi} \arcsin \sqrt{\left(\frac{a + d}{p}\right)}$	0	1	S	M	none
45	Ham[b]	Hamann (1961),[56] Holley–Guilford (1964),[57] Hubert (1977)[58]	$s_{\mathrm{Ham}} = \frac{a + d - b - c}{p}$	+1	2	S	M	none
46	McC[b]	McConaughey (1964)[59]	$s_{\mathrm{Mc}} = \frac{a^2 - bc}{(a + b)(a + c)}$	+1	2	A	N	$a = 0 \rightarrow s = 0$
47	GL[b]	Gower–Legendre (1986)[60]	$s_{\mathrm{GL}} = \frac{a + d}{a + 0.5(b + c) + d}$	0	1	S	N	none
48	BU2[b]	Baroni-Urbani–Buser (1976)[33]	$s_{\mathrm{BU2}} = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c}$	+1	2	I	M	$d = p \rightarrow s = 1$
49	Joh[b]	Johnson (1967)[61]	$s_{\mathrm{Joh}} = \frac{a}{a + b} + \frac{a}{a + c}$	0	2	A	N	$a = 0 \rightarrow s = 0$
50	Sco[b]	Scott (1955)[62]	$s_{\mathrm{Sco}} = \frac{4ad - (b + c)^2}{(2a + b + c)(2d + b + c)}$	+1	0	S	M	$a = p \vee d = p \rightarrow s = 1$
51	Maa[b]	van der Maarel (1969)[63]	$s_{\mathrm{Maa}} = \frac{2a - b - c}{2a + b + c}$	+1	2	A	N	$a = 0 \rightarrow s = 0$

[a] In the column "conditions", den indicates the denominator of the function. [b] See text.

Table 3. Activity Classes Used in the Virtual Screening Experiments: (a) MDDR, (b) WOMBAT, and (c) MUV Data Sets

(a) MDDR activity class	active molecules	active scaffolds	mean pairwise similarity
5HT3 antagonists	752	417	0.35
5HT1A agonists	827	450	0.34
5HT reuptake inhibitors	359	181	0.35
D2 antagonists	395	258	0.35
renin inhibitors	1125	554	0.57
angiotensin II AT1 antagonists	943	464	0.40
thrombin inhibitors	803	425	0.42
substance P antagonists	1246	586	0.40
HIV protease inhibitors	750	461	0.45
cyclooxygenase inhibitors	636	282	0.27
protein kinase C inhibitors	453	171	0.32

(b) WOMBAT activity class	active molecules	active scaffolds	mean pairwise similarity
5HT3 antagonists	220	117	0.38
5HT1A antagonists	592	224	0.40
D2 antagonists	910	324	0.37
renin inhibitors	474	253	0.59
angiotensin II AT1 antagonists	724	253	0.44
thrombin inhibitors	421	196	0.42
substance P antagonists	558	186	0.43
HIV protease inhibitors	1128	473	0.44
cyclooxygenase inhibitors	965	220	0.32
protein kinase C inhibitors	142	31	0.57
acetylcholine esterase inhibitors	503	220	0.37
factor Xa inhibitors	842	328	0.39
matrix metalloprotease inhibitors	694	280	0.44
phosphodiesterase inhibitors	596	270	0.36

(c) MUV activity class (assay ID)	active molecules	active scaffolds	mean pairwise similarity
sphingosine-1-phosphate 1 receptor potentiators (aid 466)	30	28	0.17
protein kinase A inhibitors (aid 548)	30	27	0.18
steroidogenic factor 1 inhibitors (aid 600)	30	24	0.18
rho kinase 2 inhibitors (aid 644)	30	27	0.18
HIV reverse transcriptase RNase (aid 652)	30	27	0.15
ephrin type-A receptor 4 antagonist inhibitors (aid 689)	30	29	0.16
steroidogenic factor 1 activators (aid 692)	30	30	0.17
heat shock protein 90 kDa alpha inhibitors (aid 712)	30	27	0.16
estrogen receptor-alpha coactivator binding inhibitors (aid 713)	30	26	0.17
estrogen receptor-beta coactivator binding inhibitors (aid 733)	30	28	0.17
estrogen receptor-alpha coactivator binding potentiators (aid 737)	30	28	0.19
focal adhesion kinase inhibitors (aid 810)	30	28	0.16
cathepsin G (aid 832)	30	24	0.22
factor XIa (aid 846)	30	21	0.22
factor XIIa (aid 852)	30	24	0.22
dopamine receptor D1 allosteric modulators (aid 858)	30	24	0.17
muscarinic receptor M1 allosteric modulators (aid 859)	30	29	0.18

the biological activity is known. There are several such data sets now available for evaluation purposes, and we have used three in the experiments reported here (so as to ensure that the results obtained are not overly dependent on the nature of the test data): the MDL Drug Data Report and World of Molecular Bioactivity databases (MDDR and WOMBAT, as described in detail by Gardiner et al.[64]) and the Maximum Unbiased Validation database (MUV, as described in detail by Rohrer and Baumann[65]).

The MDDR database (available from Accelrys Inc. at http://www.http://accelrys.com/) contains the structures and pharmacological class information for molecules that have been reported in patents, journals, and conference proceedings as exhibiting biological activity. The activity data is qualitative: a molecule is noted as exhibiting a specific activity, and it is assumed to be inactive if that is not the case. The data set used here contained 102 540 molecules, and searches were carried out for eleven classes of active compounds. The WOMBAT database (available from Sunset Molecular Discovery LLC at http://www.sunsetmolecular.com/) contains data extracted from important drug-discovery journals such as the *Journal of Medicinal Chemistry* and *Bioorganic & Medicinal Chemistry*. The activity data is quantitative but has been converted for use here to qualitative form[64] so that it is in the same form as in the other two data sets. The data set used here contained 138 127 molecules, and searches were carried out for 14 classes of active compounds. The MUV data set (available by download from http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html) is rather different in nature from the MDDR and WOMBAT data sets since it has been designed specifically for the evaluation of virtual screening systems, using structure−activity data from the PubChem database. MUV contains 17 activity classes, each containing 30 structurally diverse actives and 15 000 inactive decoy molecules. These classes have the following PubChem Assay ID (aid) codes (see http://pubchem.ncbi.nlm.nih.gov/): aid466, aid548, aid600, aid644, aid652, aid689, aid692, aid712, aid713, aid733, aid737, aid810, aid832, aid846, aid852, aid858, and aid859.

The data sets are summarized in Table 3. Each row of the table contains an activity class, the number of molecules belonging to the class, and two indications of the class's diversity (i.e., its degree of structural heterogeneity). The first diversity value in each row of the table is the number of distinct scaffolds present in the set of active molecules. The second value in each row was calculated by comparing each member of an activity class with all of the other members of that class, calculating the intermolecular similarities using the standard Unity 2D fingerprint (available from Tripos Inc. at http://www.tripos.com) and the Tanimoto coefficient, and then, computing the mean intraset similarity. These last two columns make clear the high diversity of the MUV data set. The molecules in the three data sets were represented by ECFP_4 (for Extended Connectivity Fingerprint encoding circular substructures of diameter four bonds) fingerprints from the Accelrys Pipeline Pilot software after hashing each fingerprint to a fixed length of 1024 bits.[8,66]

■ RESULTS AND DISCUSSION

Preliminary Analysis. A preliminary analysis of the similarity coefficients was performed by calculating the pairwise Pearson correlations between the coefficients for the simulated data set. This analysis identified five pairs and a triplet of coefficients that were perfectly correlated, and it was hence decided to exclude one member of each pair, as detailed in Table 4. We have chosen to retain the better known member of each pair and thus only the first 44 coefficients in Table 2 were considered further. The Appendix to this paper provides a mathematical demonstration of the existence of such perfect correlations.

Table 4. Pairs of Correlated Coefficients[a].

retained coefficients	excluded coefficients
simple matching (SM, 1)	Hamman (Ham, 45)
Kulczinski (Kul, 11)	McConaughey (McC, 46), Johnson (Joh, 49),
2nd Sneath−Sokal (SS2, 13)	Gower−Legendre (GL, 47)
Baroni-Urbani−Buser (BUB, 10)	2nd Baroni-Urbani−Buser (BU2, 48)
Rogot−Goldberg (RG, 18)	Scott (Sco, 50)
Gleason (Gle, 4)	van der Maarel (Maa, 51)

[a] The first column lists those coefficients that were retained, and the second column contains the perfectly correlated coefficients that were excluded from further analysis.

Basic Statistics and Plots. The 100 000 similarity values that had been generated for each coefficient were analyzed to calculate the following descriptive statistics: minimum and maximum values (min, max), mean, standard deviation (std), coefficient of variation (cv), 5 and 95 percentiles (perc(5) and perc(95)). These values are listed in Table 5. The minimum and maximum values of all the coefficients are 0 and 1, respectively, showing that the simulated data set had been generated so as to allow all of the different combinations of the parameters a, b, c, and d randomly chosen in the range [0, 1024], including extreme cases such as $a = p$ and $d = p$.

Inspection of Table 5 suggests that most of the coefficients have a mean value around 0.5, and that they span the similarity range in a satisfactory way. There are three very anomalous coefficients: Co1 (24), Co2 (25) yield very high values, and Mou (16) yields very low values. These outlier coefficients were probably originally proposed to deal with short vectors, where the parameters b, c, and d may have less influence than a. Less extreme behavior is exhibited by CT1 (39), CT4 (42), CT3 (41), Sim (7), SS2 (13), Di2 (32), and Kul (11) (which all have mean values greater than 0.55) and CT2 (40), SS4 (29), HL (38), and GK (27) (which all have mean values less than 0.30). Turning to the standard deviations (and excluding Co1, Co2, and Mou), the coefficients showing the maximum variability are Yu1 (20), Sor (33), Di2 (32), Di1 (31), Fos (22), For (6), Gle (4), Ja (14), BB (8), JT (3), and DK (9) (all with standard deviations greater than 0.30), while the minimum variability is provided by Den (2), CT1 (39), CT5 (43), CT2 (40), dis (26), and Coh (34) (all with standard deviations lower than 0.20).

The ordered sequences of similarity values (in ascending order) were plotted for each coefficient to explore the functional shape. In order to simplify the analysis and discussion that follows, the plots are presented in three different figures: symmetric functions (Figure 1), asymmetric functions (Figure 2), and correlation-based functions (Figure 3). Inspection of these figures shows that the shapes of the functions can be approximately categorized as logarithmic, exponential, sigmoidal, or quasi-linear in character.

We consider the symmetric coefficients in Figure 1 first. The outlier coefficients CT1 (39) and CT2 (40) exhibit markedly different behaviors, with a logarithmic shape for the former and an exponential shape for the latter. Like most of the correlation-based coefficients (see below and Figure 3), CT5 (43), SS3 (28), AC (44), and RG (18) are the only symmetric functions showing a sigmoidal behavior. Considering now the asymmetric coefficients in Figure 2, most exhibit quasi-linear behavior, increasing linearly and smoothly in the range [0, 1]. Mou (16) is clearly radically different from all the others, reflecting the very low mean value noted in Table 5. Fai (15) is approximately bilinear; CT4 (42), CT3 (41) and Sim (7) coefficients show marked logarithmic behavior; while SS1 (12), RR (5), Fos (22), For (6), JT (3), and Sor (33) show a smoothed exponential behavior. With the obvious exception of Co1 and Co2 (in the top-left corner), all of the correlation-based coefficients in Figure 3 have a sigmoidal shape. Moreover, most of them provide intermediate similarity values (around 0.5) for several different combinations of the parameters a, b, c, and d.

Multidimensional Scaling. Multidimensional scaling (MDS) is a multivariate analysis technique that enables the identification of a subspace of the original p-dimensional space into which the points can be projected and in which the interobject dissimilarities are approximated as well as possible by the corresponding interpoint distances.[67] The final result of such a technique is a geometrical model of the objects in the analysis, which allows a visual investigation of the relationships between the objects.

STATISTICA software[68] was used to carry out MDS on the 44 × 44 matrix of the pairwise Pearson correlation coefficients calculated from the simulated data. The final configuration of the binary similarity coefficients in a two-dimensional MDS plot is shown in Figure 4. This figure omits the Cole and Mountford coefficients (Co1 (24), Co2 (25), and Mou (16)) since inclusion of these significant outliers (see Figures 2 and 3) resulted in all of the other 41 coefficients in the MDS plot lying in a single tight cluster.

At a first glance, the remaining 41 similarity coefficients appear well clustered according to their symmetry properties, with the symmetric functions (green squares, on the middle-left-bottom side), the asymmetric functions (blue triangles, on the right side), and the correlation-based functions (red circles, on the left-top side) well separated from each other. In this respect, it is interesting to note that BUB (10) and Fai (15), which are intermediate in character between symmetric and asymmetric functions, are appropriately located between the symmetric and asymmetric clusters. Many of the coefficients are very near to each other in the plot, indicating close similarity relationships, e.g., the group comprising SM (1), RT (2), SS2 (13), and AC (44), which have a rank correlation equal to one. In much the same way, the group JT (3), Ja (14), Gle (4), SS1 (12), For (6), Fos (22), and DK (9) have rank correlations larger than 0.99. Some coefficients, however, are quite isolated in the MDS plot. This is the case for the pairs CT1 (39) and CT2 (40), Sim (7) and Di2 (32), CT3 (41) and *CT4* (42), and RR (5) and CT5 (43) also seem to be quite separated from other coefficients.

Rank Correlation Analysis. Spearman rank correlations were calculated for all pairs of coefficients. It was observed that some pairs of coefficients were perfectly correlated to each other, i.e., they were monotonic in providing exactly the same ranking of the cases, and near-monotonicity was observed for several other pairs. The most significant rank correlations are reported in Table 6 where, for example, row 2 shows that JT (3), Ja (14), Gle (4), and SS1 (12) are perfectly correlated with each other. In like mode, For (6), Fos (22), and DK (9) are perfectly

Table 5. Statistical Parameters for the Studied 44 Similarity Coefficients Obtained by the Simulated Data Set

no.	symbol	min	max	mean	std	cv	perc(5)	perc(95)
1	SM	0	1	0.5420	0.2869	0.5293	0.0625	0.9629
2	RT	0	1	0.4268	0.2848	0.6672	0.0323	0.9284
3	JT	0	1	0.4050	0.3112	0.7685	0.0102	0.9433
4	Gle	0	1	0.5066	0.3204	0.6325	0.0201	0.9708
5	RR	0	1	0.3329	0.2939	0.8829	0.0059	0.9014
6	For	0	1	0.3756	0.3234	0.8609	0.0015	0.9429
7	Sim	0	1	0.6902	0.2984	0.4323	0.0865	0.9971
8	BB	0	1	0.4427	0.3185	0.7194	0.0119	0.9569
9	DK	0	1	0.5302	0.3075	0.5800	0.0391	0.9710
10	BUB	0	1	0.5052	0.2980	0.5898	0.0375	0.9569
11	Kul	0	1	0.5665	0.2875	0.5075	0.0606	0.9712
12	SS1	0	1	0.3092	0.2880	0.9312	0.0051	0.8927
13	SS2	0	1	0.6532	0.2716	0.4158	0.1176	0.9811
14	Ja	0	1	0.5658	0.3187	0.5633	0.0299	0.9804
15	Fai	0	1	0.4374	0.2655	0.6069	0.0454	0.9170
16	Mou	0	1	0.0085	0.0502	5.9205	0.0001	0.0232
17	Mic	0	1	0.5099	0.2155	0.4227	0.1096	0.8912
18	RG	0	1	0.4558	0.2411	0.5290	0.0611	0.8669
19	HD	0	1	0.3619	0.2353	0.6503	0.0328	0.8031
20	Yu1	0	1	0.5349	0.3700	0.6917	0.0025	0.9978
21	Yu2	0	1	0.5252	0.2837	0.5401	0.0479	0.9551
22	Fos	0	1	0.3746	0.3234	0.8635	0.0013	0.9425
23	Den	0	1	0.3431	0.1336	0.3894	0.1270	0.5880
24	Co1	0	1	0.9921	0.0803	0.0809	0.9970	0.9999
25	Co2	0	1	0.9921	0.0802	0.0809	0.9966	0.9999
26	dis	0	1	0.5066	0.1563	0.3085	0.2244	0.7766
27	GK	0	1	0.2961	0.2579	0.8710	0.0049	0.8076
28	SS3	0	1	0.5213	0.2161	0.4146	0.1475	0.8748
29	SS4	0	1	0.2465	0.2410	0.9778	0.0018	0.7527
30	Phi	0	1	0.5175	0.2140	0.4136	0.1379	0.8727
31	Di1	0	1	0.5473	0.3248	0.5934	0.0256	0.9868
32	Di2	0	1	0.5856	0.3390	0.5788	0.0229	0.9946
33	Sor	0	1	0.4051	0.3449	0.8516	0.0007	0.9737
34	Coh	0	1	0.5378	0.1788	0.3325	0.2461	0.8673
35	Pe1	0	1	0.5210	0.2308	0.4429	0.1157	0.9014
36	Pe2	0	1	0.5167	0.2191	0.4240	0.1214	0.8899
37	MP	0	1	0.5164	0.2105	0.4076	0.1397	0.8708
38	HL	0	1	0.2687	0.2102	0.7824	0.0270	0.7124
39	CT1	0	1	0.8733	0.1354	0.1551	0.6022	0.9946
40	CT2	0	1	0.1628	0.1525	0.9364	0.0093	0.4715
41	CT3	0	1	0.7401	0.2208	0.2983	0.2807	0.9850
42	CT4	0	1	0.7734	0.2160	0.2793	0.3197	0.9914
43	CT5	0	1	0.5128	0.1400	0.2729	0.2857	0.7384
44	AC	0	1	0.5316	0.2174	0.4090	0.1609	0.8766

correlated with each other (row 3) and have correlations greater than 0.990 with JT, Ja, Gle, and SS1 (row 7).

Hasse Diagram. The Hasse diagram is an interesting technique that is used to compare rankings based on multiple criteria,[69,70] where the common ordering conditions are implemented with the relationship of incomparability. According to this theory, a coefficient A is superior to (or dominates or is "comparable with") the coefficient B if (and only if) for all the pairs of cases, the similarity value ($s_{ij}(A)$) of A is greater than (or equal to) the similarity value ($s_{ij}(B)$) of B where i and j run on all the cases:

$$\mathrm{A} \mapsto \mathrm{B} \quad \mathit{iff} \quad s_{ij}(A) \geq s_{ij}(B) \quad \forall i, j$$

When this condition is fulfilled, A is located at a higher level than B in the Hasse diagram, and they are linked together. If an inversion order appears for at least one comparison, then the coefficients A and B are not considered comparable: they are hence located at the same level in the Hasse diagram and do not share a link. A coefficient that is not comparable with any other coefficient is conventionally located in the highest level. Each path from the highest to the lowest levels provides a complete order, meaning that all the similarity values of a coefficient at higher level are greater than (or equal to) the similarity values of coefficients linked at lower levels.

The Hasse diagram was calculated using the DART[71] software in a version that allowed the processing of the 100 000 cases comprising the simulated data set. The resulting Hasse diagram is shown in Figure 5, where asymmetric coefficients are represented in blue, symmetric coefficients in green, and correlation-based coefficients in red. The correlation-based coefficients give, on average, the largest similarity values, followed by the symmetric coefficients. The asymmetric indices are spread throughout all seven levels of the diagram; however, all the asymmetric indices at levels one and two (JT, RR, For, and SS1) have an exponential

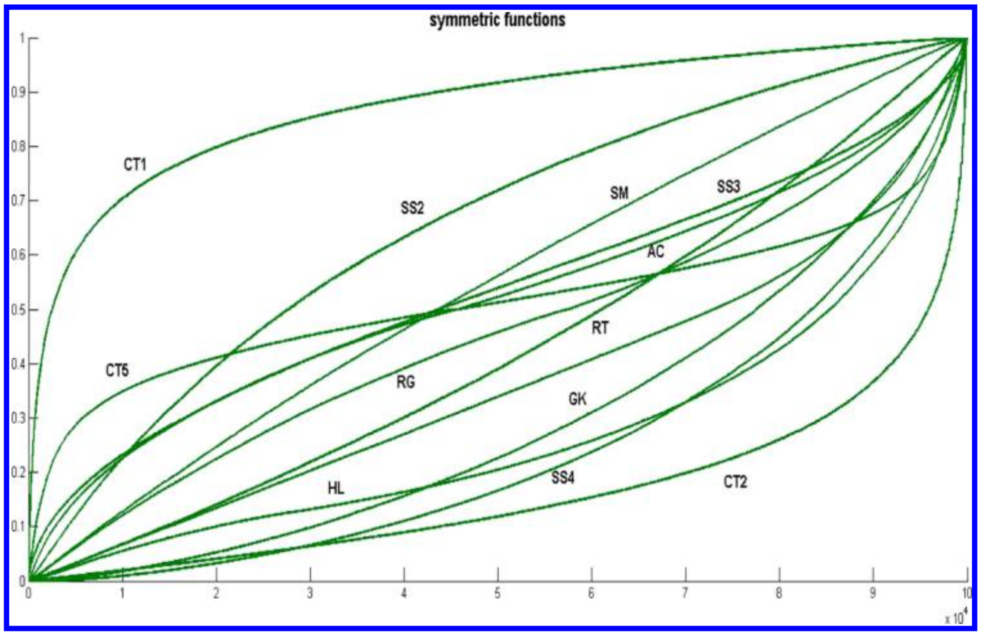


Figure 1. Line plots of the symmetric binary coefficients calculated from the simulated data set.

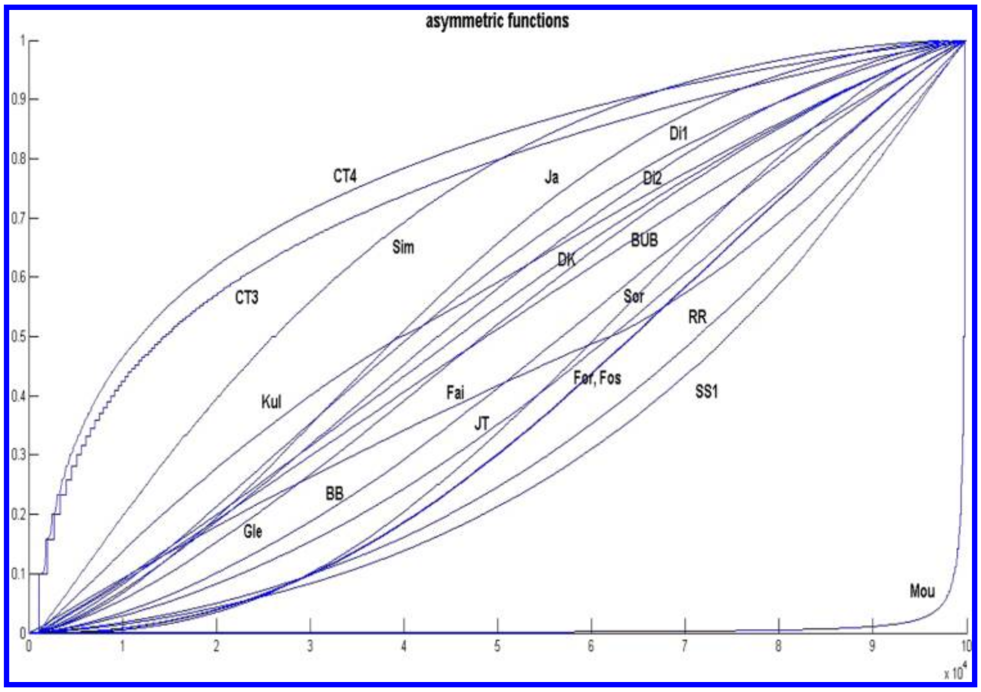


Figure 2. Line plots of the asymmetric binary coefficients calculated from the simulated data set.

shape that tends to provide low similarity values for most of the cases. These coefficients are all influenced by the size of the binary string, as already observed by Holliday et al.[72] for the Jaccard–Tanimoto coefficient (JT, 3).

There are just three coefficients that are not comparable with any of the others, and that are hence located at level 7 without links, viz Mic (17), Fos (22), and CT5 (43) coefficients. Simple one-dimensional paths are the following:

$$0 \leq \text{GK} \leq \{\text{Yu2, Phi}\} \leq 1,\ 0 \leq \text{SS4} \leq \{\text{Yu1, Yu2, Phi}\} \leq 1$$

$$0 \leq \{\text{Pe1, Pe2, MP}\} \leq \{\text{Co1, Co2}\} \leq 1$$

$$0 \leq \text{Mou} \leq \text{Sim} \leq 1$$

Inspection of the diagram shows that the similarity values of Yu2 (21) are always greater than (or equal to) the values of GK (27) and that the values of the two coefficients Yu1 (20) and Yu2 (21) together with Phi (43) are always greater than (or equal to) the values of SS4 (29). In like manner, the values of the two coefficients Co1 (24) and Co2 (25) are always greater than (or equal to) the values of Pe1 (35), Pe2 (36), and MP (37); and the same consideration holds for Sim (7) with respect to Mou (16). The Mou (16), Den (23), Pe1 (35), Pe2 (36), and MP (37) coefficients are not comparable with other coefficients already at level 6, Sor (33) is no longer comparable at level 5, while GK (27), SS4 (29), HL (38), and CT2 (40) become no longer comparable at lower levels.

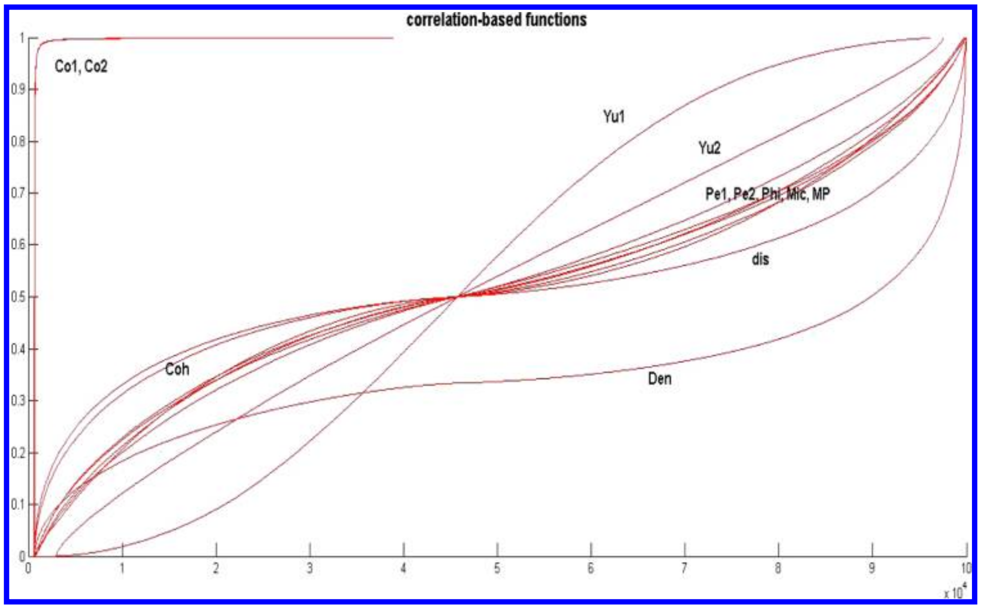


Figure 3. Line plots of the correlation-based binary coefficients calculated from the simulated data set.

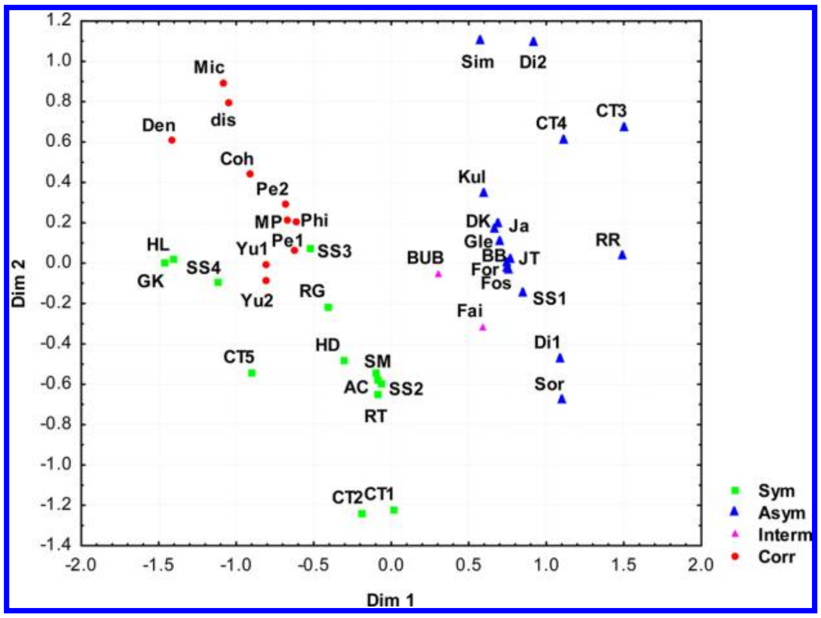


Figure 4. Multidimensional scaling of the binary similarity coefficients. The Co1, Co2, and Mou coefficients were excluded from analysis as they are strong outliers (see text).

A path running over all the seven levels of the Hasse diagram is

$$0 \leq \{\text{RR, For, SS1}\} \leq \text{JT} \leq \text{BB} \leq \text{Gle} \leq \text{DK} \leq \text{Kul} \leq \text{Sim} \leq 1$$

The sequence from Sim to JT confirms what had already been demonstrated in the work of Warrens.[17] The same sequence is also present in the ATDM plot (see Figure 6 below). Moreover, as already observed by other authors,[72] this sequence confirms that the similarity values of the Jaccard–Tanimoto coefficient (JT, 3) are lower similarity estimates than those provided by several other coefficients. Finally, other paths emerging from the Hasse diagram are

$$0 \leq \text{SS1} \leq \text{RT} \leq \text{SM} \leq \text{SS2} \leq \text{CT1} \leq 1$$

$$0 \leq \{\text{SS1, SS4, HL, CT2}\} \leq \text{RT} \leq \text{SM} \leq \text{SS2} \leq \text{CT1} \leq 1$$

$$0 \leq \text{RR} \leq \text{CT3} \leq \text{CT4} \leq \text{CT1} \leq 1$$

In all of these cases, the first Consonni–Todeschini coefficient (CT1, 39) appears as an upper bound to the other similarity measures.

Atemporal Target Diffusion Model. ATDM is a recently proposed algorithm[73] that has been developed to detect the dependencies among pairs of variables in a large data set, while also taking approximate account of their higher order relationships with other variables. The theoretical basis of ATDM is explained briefly below.

Let n be the number of samples, p be the number of variables, and x be a variable such that $x \in [0, 1]$. The new relationship between two variables, which is interpreted as a connection strength, is calculated from

Table 6. Most Significant Rank Correlations between the 44 Binary Similarity Coefficients, Provided by the Simulated Data Set

row	spearman	binary similarity coefficients
1	$\rho = 1$	SM, RT, SS2, AC, CT1, CT2
2		JT, Ja, Gle, SS1
3		DK, For, Fos
4		Yu1, Yu2
5		Di1, Sor
6	$\rho \geq 0.990$	Coh, MP, Mic, Den, dis
7		JT, Ja, Gle, SS1, For, DK, Fos
8	$\rho \geq 0.980$	Coh, MP, Mic, Den, dis, Pe1, Pe2
9		HL, SS4
10	$\rho \geq 0.970$	CT5, SS3
11		Yu1, Yu2, Phi
12		JT, Ja, Gle, SS1, For, DK, Fos, CT4

$$c_{jk} = \sum_{i=1}^{n} \left[x_{ij} x_{ik} \sum_{\substack{m \neq j \\ m \neq k}}^{p} \frac{(1+\varepsilon) + (x_{ij} - x_{im})}{(1+\varepsilon) + (x_{ik} - x_{im})} \right]$$

where i runs over all the n samples, j and k denote the jth/kth pair of variables, and m runs over all the $p - 2$ variables other than j and k. The parameter ε (which was set to be 0.0001 here) is included to avoid singularities. The connection w_{jk} between two variables is then calculated from

$$w_{jk} = c_{jk} e^{-d_{jk}/\sqrt{\alpha}}$$

where d_{jk} is any distance measure. The contribution made by the distance is determined by the variable weighting parameter α, which was set to 0.1 in the work reported here. The distance used in this work is the average Euclidean distance:

$$d_{jk} = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - x_{ik})^2}{n}}$$

where x_{ij} and x_{ik} are the similarity values provided by the jth and kth coefficients for the ith case.

Different limit cases are relevant for the analysis of this function:

1. When the intersection between the variables x_{ij} and x_{ik} is null: in this case, the value of the function is zero; this means that in the record (i) the two variables are completely independent, because they do not share information;
2. When the intersection between the variables x_{ij} and x_{ik} is not null, four different limit subcases need to be distinguished:
   (a) If all the other variables, x_{im}, are null, then the connection strength will be given by the intersection of the two considered variables, x_{ij} and x_{ik}, weighted by their ratio.
   (b) The numerator and the denominator of the summation (m) are equal: in this case, because the variables have the same value, the strength of their connection is their intersection.
   (c) If the numerator of the summation (m) is bigger than the denominator, then the connection strength will have a high value; this solution implies that because x_{ij} is systematically bigger than x_{ik} and x_{im}, then x_{ij} is an outlayer and x_{ik} is or an average value or x_{ik} is a very small value.
   (d) If the denominator of the summation (m) is bigger than the nominator, then the connection strength will have a low value; this solution implies that because x_{ik} is systematically bigger than x_{ij} and x_{im}, then x_{ik} is an outlyer and x_{ij} is or an average value or x_{ij} is a very small value.

In practice, ATDM equations weight how much each variable depends on any other, while also considering the contexts of the other variables. Then, we can say that these equations weight the dependency of any pair of variables with an approximation of the highest order of relationship. Consequently, the connection w_{jk} is a coefficient of *similarity and causality* based on the two variables' intersection and weighted by both the Euclidean distance and the values of the other variables $\mathbf{x}_m$. The contribution of the other variables (m) makes the "cause−effect" relationship between j and k bigger if j tends to be an outlier and k has a typically average value; conversely, the cause−effect relationship will be smaller in the opposite case. The $\mathbf{W}$ matrix is symmetric when all the x values are 1 or 0, while it is asymmetric in all other cases.

The ATDM algorithm was applied to the large simulated data set by means of the ATDM software developed by Buscema.[73] The results of an ATDM analysis are given in terms of a minimum spanning tree (MST) calculated on the $\mathbf{W}$ matrix of the connections w_{jk} between all the possible pairs of similarity

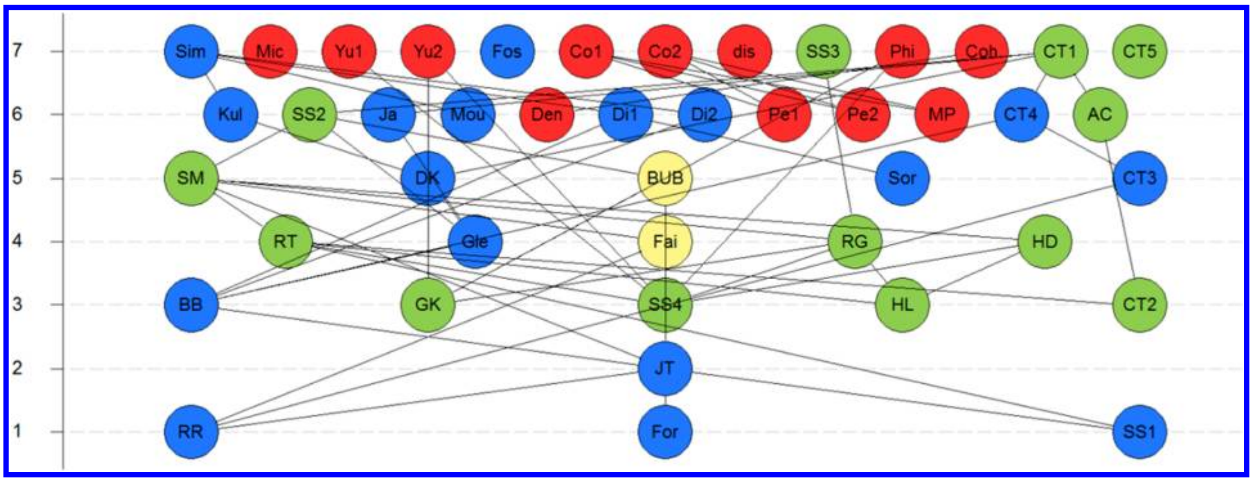


Figure 5. Hasse diagram of the binary similarity coefficients.

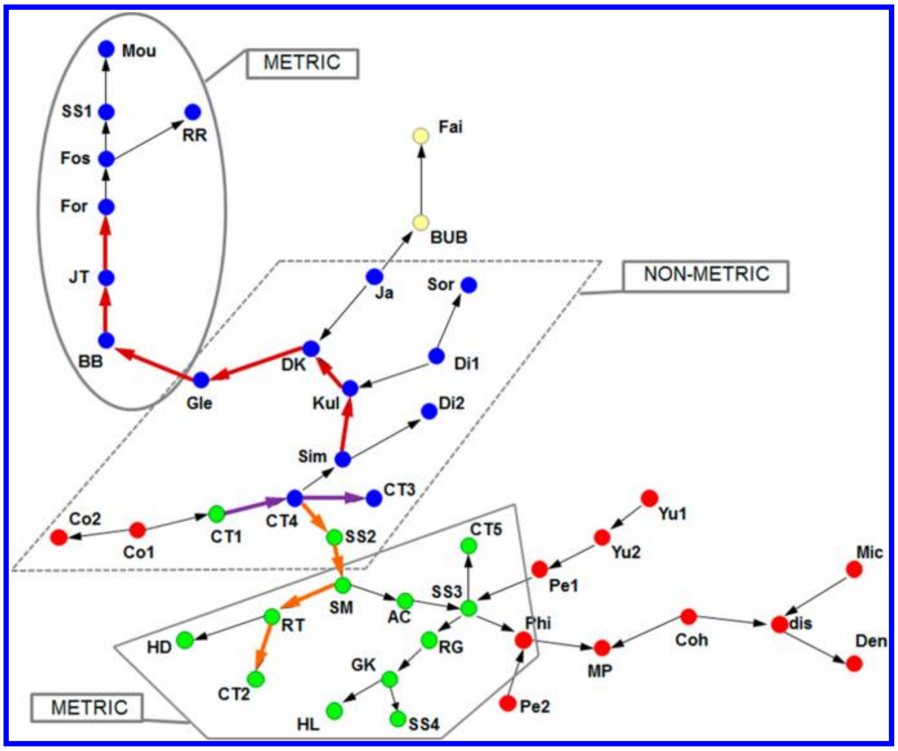


Figure 6. Minimum spanning tree calculated by the ATDM method. The frames represent coefficients having the same metricity, and bold-faced paths represent paths obtained from the Hasse diagram.

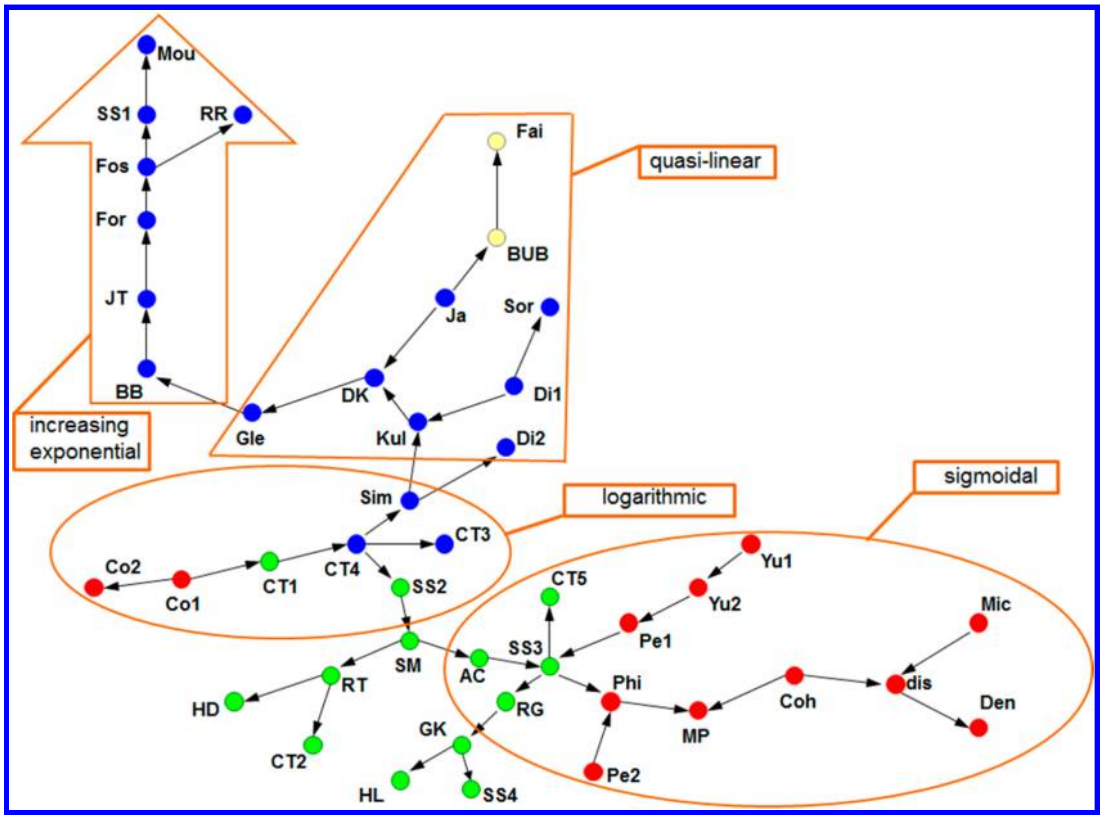


Figure 7. Minimum spanning tree calculated by the ATDM method. The frames represent coefficients having similar functional shapes.

coefficients, as shown in Figures 6 and 7. In these figures, the nodes representing the similarity coefficients are colored depending on whether they are symmetric (green), asymmetric (blue), or correlation-based (red); the two intermediate coefficients are colored in yellow.

Note that the MST is a *directed* graph when the connection matrix **W** is asymmetric. That seems to be hard to understand, because in graph theory the MST is an undirected connected graph by definition. Consequently it is impossible to generate the MST from the asymmetric weights matrix. But a *directed*

MST should be a suitable way to represent simultaneously the energy minimization criterion and the clustering constrains among a set of independent variables. To avoid the theoretical limitation imposed by the classic definition of MST, we worked in this way:

1. We start with an asymmetric weights matrix ($p \times p$) provided by the ATDM algorithm.
2. We transform the original asymmetric weights matrix into an "asymmetric distance matrix".
3. We calculate the average between each pair of variables in the asymmetric distance matrix.
4. We calculate from the new symmetric distance matrix the $p - 1$ edges for a regular MST.
5. We consider again the asymmetric distance matrix, and we calculate the shorter distance for each pair of variables directly connected in the regular MST.
6. We assume that the shortest length between a pair of variables indicates the direction of the connection.
7. Consequently any edge is transformed into an arc in the new "directed MST".

The MDS plot in Figure 4 mainly contains information about the separation of symmetric, asymmetric, and correlation-based coefficients, whereas the ATDM approach allows the identification of further patterns and relationships among the similarity indices. In particular, in Figure 6, clusters of metric and nonmetric coefficients are well-defined for most of the coefficients; and some of the paths represented by bold arrows reflect interesting paths observed in the Hasse diagram (Figure 5). In Figure 7, the clusters that are present can readily be identified with the different functional shapes of the coefficients as noted in Figures 1−3.

The ATDM plots in Figures 6 and 7 provide a much richer characterization of the relationships between the coefficients than those that can be deduced from the MDS plot in Figure 4. For example, CT1 (39) and CT2 (40) are both metric functions but have opposite shapes in Figure 1: in the MDS plot, they appear very similar, but in the ATDM graph they belong to different clusters and are separated by a path of length 5.

Analogously, CT5 (43), AC (44), and RG (18) are well-separated in the MDS plot; in ATDM, however, they appear more similar to each other and to the correlation-based functions, all displaying an obvious sigmoidal shape. Information about the shape (see Figures 1 and 2) is also correctly taken into account in ATDM for CT1 (39), CT3 (41), and CT4 (42), which are all directly linked together, whereas CT1 (39) is considered very different from the other two coefficients in MDS. The coefficients of Jaccard−Tanimoto (JT, 3) and Jaccard (Ja, 14) appear very similar in MDS, but are significantly separated (by a path of length 4) in ATDM. Their shapes in Figure 2 are quite different, and the path linking the two coefficients (Ja−DK−Gle−BB−JT) seems to reflect very well the transition in metricity (the first three are nonmetric, while the last two are metric), their symmetry (all asymmetric), and the transition in shape. Again, the similarity of SS3 (28) with CT5 (43) and AC (44) is not as marked in the MDS plot as it is in the ATDM plot: in effect, they are all metric and symmetric functions and their shape (Figure 1−3) appears quite similar. An opposite consideration holds for the similarity of SS1 (12, asymmetric and metric) and Di1 (31, asymmetric and nonmetric), which appear to be very similar in the MDS plot but are quite different in the ATDM plot, the latter mirroring the difference in shapes that is seen in Figure 2.

Virtual Screening on Chemical Data Sets. There are many ways in which the effectiveness of a virtual screening system can be evaluated. Here, we have used probably the simplest criterion, which is the recall, i.e., the number of active molecules having the same activity as the reference structure and occurring at or above some cutoff position in the ranking resulting from a similarity search. Specifically, each search was evaluated by the number of such active molecules in the top-1% of a ranking. For the searches of the MDDR and WOMBAT data sets, ten actives were chosen at random from each activity class to act as the reference structures, while all of the 30 actives for each class were used as reference structures for the searches of the MUV data set. The results for each of the similarity coefficients were averaged by taking the median number of actives retrieved using that coefficient for the ten (or 30) searches in each activity class, and the 44 coefficients were then ranked in order of decreasing median numbers of actives for each activity class.

A similarity coefficient needs to provide effective search performance across the full range of types of bioactivity if it is to be of general applicability as a virtual screening tool. To test whether this is indeed the case, the results were evaluated using Kendall's coefficient of concordance (W), which is used to evaluate the consistency of k different sets of ranked judgements of the same set of N different objects.[74] In the present context, each of the activity classes ranks the 44 different coefficients in order of decreasing effectiveness (as measured by the median number of actives retrieved); thus, $N = 44$ and $k = 11$, 14, or 17 (for the MDDR, WOMBAT, or MUV data sets, respectively). The statistical significance of the computed value of W can be tested using the χ^2 distribution since $\chi^2 = k(N - 1)W$ with $N - 1$ degrees of freedom (if $N > 7$ as is the case here). If a significant value is obtained, then Siegel and Castellan suggest that the best overall ranking of the N objects can be obtained using their mean ranks averaged over the k judges.[74]

Statistically significant levels of concordance across the activity classes were observed for both the MDDR and WOMBAT data sets as shown in Table 7. This being so, it is reasonable to generate overall rankings for MDDR and WOMBAT data sets (but not for the MUV results where $0.05 \leq p \leq 0.1$), as shown in Table 8 (where the lower the mean rank the greater the ability of a coefficient to identify actives in a similarity search). Inspection of the MDDR and WOMBAT columns in the table demonstrates a very high degree of resemblance throughout the entire ranked list. For example, coefficients JT (3), Gle (4), SS1 (12), Ja (14) were ranked first equal over the eleven activity classes comprising the MDDR data set, with coefficient CT4 (42) ranked fifth and numerically very near to the first block. This block includes the Tanimoto coefficient (JT), which is the coefficient of choice in most operational similarity searching systems.[8] CT4 and HL (38) also have high ranks, and these six coefficients also occur right at the top of the WOMBAT ranking.

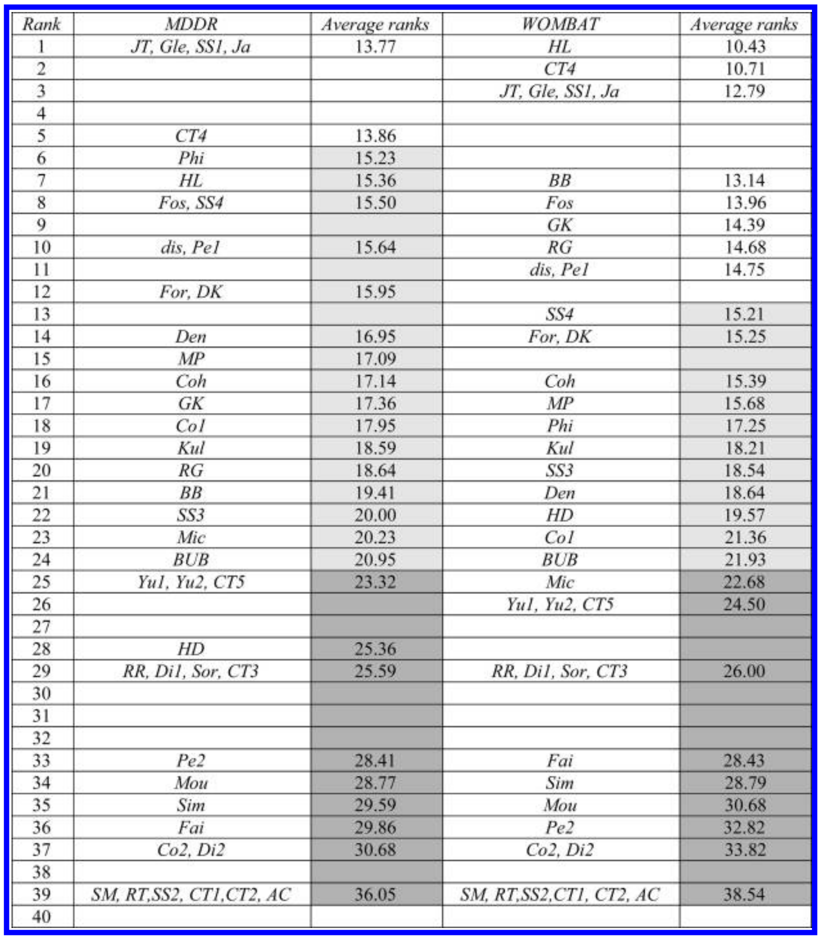
Table 7. Kendall's Test of Concordance Results

statistic	MDDR	WOMBAT	MUV
W	0.343	0.504	0.181
χ^2	162.748	303.47	132.386
p	$p \leq 0.05$	$p \leq 0.005$	$p \leq 0.1$

Taking the mean of the two ranks for each coefficient in Table 8, the first three positions in the overall ranking are

$$\text{CT4}(12.29) < \text{HL}(12.90) < \text{JT, Gle, SS1, Ja}(13.28)$$

Table 8. Rank Positions of Each of the 44 Coefficients When Averaged over All of the Activity Classes for Each of the Three Data Sets[a]



Rank	MDDR	Average ranks	WOMBAT	Average ranks
1	JT, Gle, SS1, Ja	13.77	HL	10.43
2			CT4	10.71
3			JT, Gle, SS1, Ja	12.79
4				
5	CT4	13.86		
6	Phi	15.23		
7	HL	15.36	BB	13.14
8	Fos, SS4	15.50	Fos	13.96
9			GK	14.39
10	dis, Pe1	15.64	RG	14.68
11			dis, Pe1	14.75
12	For, DK	15.95		
13			SS4	15.21
14	Den	16.95	For, DK	15.25
15	MP	17.09		
16	Coh	17.14	Coh	15.39
17	GK	17.36	MP	15.68
18	Co1	17.95	Phi	17.25
19	Kul	18.59	Kul	18.21
20	RG	18.64	SS3	18.54
21	BB	19.41	Den	18.64
22	SS3	20.00	HD	19.57
23	Mic	20.23	Co1	21.36
24	BUB	20.95	BUB	21.93
25	Yu1, Yu2, CT5	23.32	Mic	22.68
26			Yu1, Yu2, CT5	24.50
27				
28	HD	25.36		
29	RR, Di1, Sor, CT3	25.59	RR, Di1, Sor, CT3	26.00
30				
31				
32				
33	Pe2	28.41	Fai	28.43
34	Mou	28.77	Sim	28.79
35	Sim	29.59	Mou	30.68
36	Fai	29.86	Pe2	32.82
37	Co2, Di2	30.68	Co2, Di2	33.82
38				
39	SM, RT,SS2, CT1,CT2, AC	36.05	SM, RT,SS2,CT1, CT2, AC	38.54
40				

[a] Backgrounds: blank (average ranks 1−15), light grey (average ranks 15−22), dark grey (average ranks > 22).

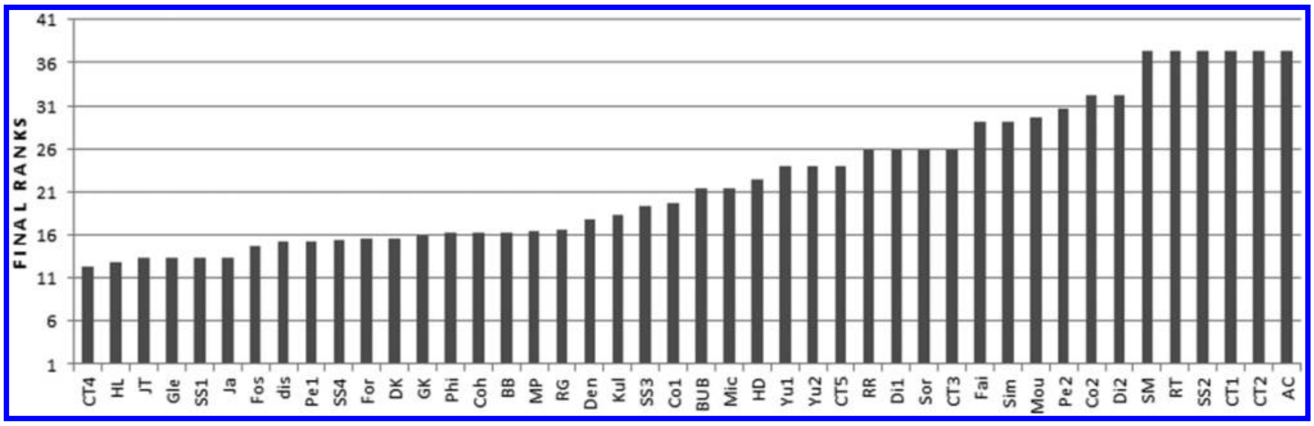


Figure 8. Global performance of the 44 binary coefficients, as obtained from the MDDR and WOMBAT data sets, ordered from best (low mean rank values) to worst (high mean rank values).

The final average ranks for all the binary similarity coefficients are shown in the histogram of Figure 8, where the six coefficients above are grouped at the left-hand of the figure with an upturn in mean rank before the start of the remaining 38 coefficients.

■ CONCLUSIONS

In this paper, 51 binary similarity coefficients (or 44 after elimination of seven redundant coefficients) were analyzed from several different points of view in order to perform an extensive comparison of their properties, both for their use in general (as evaluated by a range of statistical approaches) and for their use specifically for chemical similarity searching (as evaluated by simulated virtual screening searches using MDDR, WOMBAT, and MUV data). The novel atemporal target diffusion model approach seems able to provide a large amount of information which could not be retrieved by the common multivariate methods, in particular its ability to take account of the many nonlinear relationships among the cohort of similarity coefficients.

A group of six coefficients stands out from the remainder in the virtual screening experiments: CT4, HL, JT, Gle, SS1, and Ja. This includes the widely used Jaccard–Tanimoto coefficient but also several others that may warrant further study. Inspection of Table 2 will show that three of them—Gle, SS1, and Ja—have very similar formulations to JT, differing only in the weightings of the *a*, *b*, and *c* components of the overall coefficient. Indeed, these four coefficients yield perfectly correlated rankings, as demonstrated in Table 6. HL and CT4 have very different formulas, both from each other and from the group of four. Of the six, HL is the only symmetric coefficient, thus tending to confirm the view that it is generally more important in chemoinformatic applications to take account of the features that are present in one or both vectors than it is to take account of those that are absent from both. JT and SS1 are the only metric coefficients from among the six, suggesting that metricity is not a necessary requirement for effective screening, while the MSTs computed using the ADTM approach in Figure 7 show that the six encompass all four types of shape behavior. In conclusion, this study has demonstrated the considerable merits of the well-established Jaccard–Tanimoto coefficient; however, it has additionally identified two further coefficients that may be worthy of future study for applications in chemoinformatics.

APPENDIX

Theoretical Correlations between Binary Similarity Coefficients

Some similarity measures are theoretically correlated between them; this appendix demonstrates these correlations, some of which were already presented in the Ph.D. thesis of M. J. Warrens.[17]

Sokal–Michener (or simple matching, SM, 1) and Hamann (Ham, 45) coefficients are perfectly correlated as can be easily demonstrated by rescaling the Hamann similarity between zero and one:

$$\mathrm{Ham} = \frac{\left[\left(\frac{a+d-b-c}{p}\right)+1\right]}{2} = \frac{a+d-b-c+p}{2p}$$
$$= \frac{2a+2d}{2p} = \frac{a+d}{p} = \mathrm{SM}$$

Analogously, it can be shown that McConaughey (McC, 46) and Kulczynski (Kul, 11) coefficients are correlated by rescaling McC between zero and one.

$$\mathrm{McC} = \frac{\frac{a^2-bc}{(a+b)(a+c)}+1}{2} = \frac{a^2-bc}{2(a+b)(a+c)} + \frac{1}{2}$$
$$= \frac{1}{2}\left[\frac{a^2-bc+(a+b)(a+c)}{(a+b)(a+c)}\right]$$
$$= \frac{1}{2}\cdot\left[\frac{a^2+ac+a^2+ab}{(a+b)(a+c)}\right]$$
$$= \frac{1}{2}\left[\frac{a(a+c)+a(a+b)}{(a+b)(a+c)}\right]$$
$$= \frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right] = \mathrm{Kul}$$

A similar demonstration can be made for the van der Maarel (Maa, 51) and Gleason (Gle, 4) coefficients:

$$\mathrm{Maa} = \frac{\left(\frac{2a-b-c}{2a+b+c}\right)+1}{2} = \frac{1}{2}\left[\frac{2a-b-c+2a+b+c}{2a+b+c}\right]$$
$$= \frac{1}{2}\left[\frac{4a}{2a+b+c}\right] = \frac{2a}{2a+b+c} = \mathrm{Gle}$$

Moreover, the Johnson coefficient (Joh, 49) is trivially twice the Kul coefficient, i.e. Joh = 2Kul; and the second Sokal–Sneath coefficient (SS2, 13) is twice the Gower–Legendre coefficient (GL, 47), i.e. SS2 = 2GL. Finally, the two coefficients of Baroni-Urbani and Buser BUB (10) and BUB2 (48) are correlated with each other:

$$\mathrm{BUB2} = \frac{\left[\frac{\sqrt{ad}+a-b-c}{\sqrt{ad}+a+b+c}\right]+1}{2}$$
$$= \frac{1}{2}\left[\frac{\sqrt{ad}+a-b-c+\sqrt{ad}+a+b+c}{\sqrt{ad}+a+b+c}\right]$$
$$= \frac{1}{2}\left[\frac{2\cdot\sqrt{ad}+2a}{\sqrt{ad}+a+b+c}\right]$$
$$= \frac{\sqrt{ad}+a}{\sqrt{ad}+a+b+c} = \mathrm{BUB}$$

AUTHOR INFORMATION

Corresponding Author

*Tel.: +39-0264482820. Fax: +39-0264482839. E-mail address: roberto.todeschini@unimib.it.

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Alvarez J.; Stoichet B. *Virtual Screening in Drug Discovery*; CRC Press: Boca Raton (FL), 2005.

(2) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(3) Rippenhausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.

(4) Schneider, G. Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **2010**, *9*, 273–276.

(5) Johnson M. A.; Maggiora G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York (NY), 1990.

(6) Willett, P. Similarity methods in chemoinformatics. *Ann. Rev. Inform. Sci. Technol.* **2009**, *43*, 3–71.

(7) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal components analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(8) Leach A. R.; Gillet V. J. *An Introduction to Chemoinformatics*; Kluwer: Dordrecht, The Netherlands, 2007.

(9) Sheridan, R. P. Chemical similarity searches: when is complexity justified? *Expert Opin. Drug Discov.* **2007**, *2*, 423–430.

(10) Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Mod.* **2010**, *29*, 157–170.

(11) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D

fragment bit-strings. *Combin. Chem. High-Throughput Screening* **2002**, *5*, 155−166.

(12) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771−748.

(13) Batagelj, V.; Bren, M. Comparing resemblance measures. *J. Classif.* **1995**, *12*, 73−90.

(14) Cheetham, A. H.; Hazel, J. E. Binary (presence-absence) similarity coefficients. *J. Paleontol.* **1969**, *43*, 1130−1136.

(15) Hubalek, Z. Coefficients of Association and Similarity, based on Binary (Presence-Absence) data: An Evaluation. *Biol. Rev.* **1982**, *57*, 669−689.

(16) Legendre P.; Legendre L. *Numerical Ecology*; Elsevier: Amsterdam, The Netherlands, 1998.

(17) Warrens, M. J. *Similarity Coefficients for Binary Data*. Ph.D. Thesis, Leiden University, 2008; p 253.

(18) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdisc. Rev.: Comput. Molec. Sci.* **2011**, *1*, 260−282.

(19) Willett, P. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint-based measures of molecular resemblance. *WIRES Data Mining Knowledge Disc.* **2011**, *1*, 241−251.

(20) Sokal, R. R.; Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas. Sci. Bull.* **1958**, *38*, 1409−1438.

(21) Rand, W. Objective criteria for the evaluation of clustering methods. *J. Amer. Statist. Assoc.* **1971**, *66*, 846−850.

(22) Rogers, D. J.; Tanimoto, T. T. A computer program for classifying plants. *Science* **1960**, *132*, 1115−1118.

(23) Jaccard, P. The distribution of the flora of the alpine zone. *New Phytol.* **1912**, *11*, 37−50.

(24) Gleason, H. A. Some applications of the quadrat method. *Bull. Torrey Botanical Club* **1920**, *47*, 21−33.

(25) Dice, L. R. Measures of the amount of ecological association between species. *Ecology* **1945**, *26*, 297−302.

(26) Sørenson, T. A method for establishing groups of equal amplitude in plant sociology based on similarity of species content. *Biologiske Skrifter* **1948**, *5*, 1−34.

(27) Russell, P. F.; Rao, T. R. On habitat and association of species of Anopheline larvae in South.Eastern Madras. *J. Malaria Inst. India* **1940**, *3*, 153−178.

(28) Forbes, S. A. On the local distribution of certain Illinois fishes: An essay in statistical ecology. *Bull. Illinois State Lab. Nat. History* **1907**, *7*, 273−303.

(29) Simpson, G. G. Mammals and the nature of continents. *Am. J. Sci.* **1943**, *241*, 1−31.

(30) Braun-Blanquet J. *Plant Sociology: The Study of Plant Communities*; McGraw-Hill: New York, 1932.

(31) Driver, H. E.; Kroeber, A. L. Quantitative expression of cultural relationship. *Univ. California Publ. Am. Archaeol. Ethnol.* **1932**, *31*, 211−256.

(32) Ochiai, A. Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bull. Jpn. Soc. Fish Sci.* **1957**, *22*, 526−530.

(33) Baroni-Urbani, C.; Buser, M. W. Similarity of binary data. *Syst. Zool.* **1976**, *25*, 251−259.

(34) Kulczynski, S. Die Pflanzenassociationen der Pienenen. *Bull. Int. L'Acad. Polonaise Sci. Lett., Classe Sci. Math. Nat., Ser. B, Suppl. II* **1927**, *2*, 57−203.

(35) Sokal R. R.; Sneath P. H. A. *Principles of numerical taxonomy*; W.H. Freeman: San Francisco, CA, 1963.

(36) Faith, D. P.; Minchin, P. R.; Belcin, L. Compositional dissimilarity as a robust measure of ecological distance. *Plant Ecol.* **1987**, *69*, 57−68.

(37) Mountford M. D. An index of similarity and its applications to classificatory problems. In *Progress in Soil Zoology*; Murphy P. W., Ed.; Butterworths: London (UK), 1962; pp 43−50.

(38) Michael, E. L. Marine ecology and the coefficient of association. *J. Animal Ecol.* **1920**, *8*, 54−59.

(39) Rogot, E.; Goldberg, I. D. A proposed index for measuring agreement in test-retest studies. *J. Chronic Disease* **1966**, *19*, 991−1006.

(40) Hawkins R. P.; Dotson V. A. Reliability scores that delude: An Alice in Wonderful trip through the misleading characteristics of interobserver agreement scores in interval coding. In *Behavior Analysis: Areas of Research and Application*; Ramp, E., Semb, G., Eds.; Prentic-Hall: Englewood Cliffs, NJ, 1968.

(41) Yule, G. U. On the association of attributes in statistics. *Philos. Trans. R. Soc. A* **1900**, *75*, 257−319.

(42) Yule, G. U. On the methods of measuring association between two attributes. *J. R. Stat. Soc.* **1912**, *75*, 579−642.

(43) Cole, L. C. The measurement of interspecific association. *Ecology* **1949**, *30*, 411−424.

(44) Choi, S.-S.; Cha, S.-H.; Tappert, C. C. A Survey of Binary Similarity and Distance Measures. *J. Syst., Cyber. and Inform.* **2012**, *8*, 43−48.

(45) Goodman, L. A.; Kruskal, W. H. Measures of association for cross classifications. *J. Amer. Stat. Assoc.* **1954**, *49*, 732−764.

(46) Pearson, K.; Heron, D. On theories of association. *Biometrika* **1913**, *9*, 159−315.

(47) Wallace, D. L. A method for comparing two hierarchical clusterings: Comment. *J. Am. Stat. Assoc.* **1983**, *78*, 569−576.

(48) Post, W. J.; Snijders, T. A. B. Nonparametric unfolding models for dichotomous data. *Methodika* **1993**, *7*, 130−156.

(49) Sorgenfrei, T. Molluscan assemblages from the marine middle Miocene of South Jutland and their environments. *Danmark Geologiske Undersøgelse. Serie 2* **1959**, *79*, 403−408.

(50) Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Measurements* **1960**, *20*, 37−46.

(51) Peirce, C. S. The numerical measure of the success of predictions. *Science* **1884**, *4*, 453−454.

(52) Maxwell, A. E.; Pilliner, A. E. G. Deriving coefficients of reliability and agreement for ratings. *Brit. J. Math. Stat. Psychol.* **1968**, *21*, 105−116.

(53) Harris, F. C.; Lahey, B. B. A method for combining occurrence and nonoccurrence agreement scores. *J. Appl. Behav. Anal.* **1978**, *11*, 523−527.

(54) Consonni, V.; Todeschini, R. New similarity coefficients for binary data. *MATCH Commun. Math. Comput. Chem.* **2012**, *68*, 581−592.

(55) Austin, B.; Colwell, R. R. Evaluation of Some Coefficients for Use in Numerical Taxonomy of Microorganisms. *Int. J. Syst. Bacteriol.* **1977**, *27*, 204−210.

(56) Hamann, U. Merkmalsbestand und Verwandtschaftsbezienhungen der Farinose. Ein Betrag zum System der Monokotyledonen. *Willdenowia* **1961**, *2*, 639−768.

(57) Holley, J. W.; Guilford, J. P. A note on the G-index of agreement. *Educ. Psychol. Measurement* **1964**, *24*, 749−753.

(58) Hubert, L. J. Nominal scale response agreement as a generalized correlation. *Brit. J. Math. Stat. Psych.* **1977**, *30*, 98−103.

(59) McConnaughey, B. H. The determination and analysis of plankton communities. *Mar. Res.* **1964**, No. Special No., Indonesia, 1−40.

(60) Gower, J. C.; Legendre, P. Metric and Euclidean properties of dis-similarity coefficients. *J. Classification* **1986**, *3*, 5−48.

(61) Johnson, S. C. Hierarchical clustering schemes. *Psychometrika* **1967**, *32*, 241−254.

(62) Scott, W. A. Reliability of content analysis:The case of nominal scale coding. *Public Opin. Q.* **1955**, *19*, 321−325.

(63) van der Maarel, E. On the use of ordination models in phytosociology. *Vegetatio* **1969**, *19*, 21−46.

(64) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N. Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance. *Stat. Anal. Data Mining* **2009**, *2*, 103−114.

(65) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169−184.

(66) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(67) Krzanowski W. J. *Principles of Multivariate Analysis*; Oxford Univ. Press: New York, 1988.

(68) *STATISTICA*, ver. 7.1. StatSoft, Padova, Italy.

(69) Brüggemann, R.; Bartel, H.-G. A Theoretical Concept to Rank Environmentally Significant Chemicals. *J. Chem. Inf. Comp. Sci.* **1999**, *39*, 211−217.

(70) Brüggemann, R.; Bücherl, C.; Pudenz, S.; Steinberg, E. W. Application of the Concept of Partial Order on Comparative Evaluation of Environmental Chemicals. *Acta Hydrochim. Hydrobiol.* **1999**, *27*, 170−178.

(71) DART software (Decision Analysis by Ranking Techniques), 2007; www.talete.mi.it.

(72) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.

(73) Buscema, M. ATDM Algorithm. In *Modular Auto Associative ANNs*, version 16.0; Semeion Software no. 51, Rome, Italy, 2008−2012; www.semeion.it.

(74) Siegel S.; Castellan, N. J. *Nonparametric Statistics for the Behavioural Sciences*, second ed.; McGraw-Hill: New York, 1988.