Distance Phenomena in High-Dimensional Chemical Descriptor Spaces: Consequences for Similarity-Based Approaches

MATTHIAS RUPP,[1] PETRA SCHNEIDER,[2] GISBERT SCHNEIDER[1]
[1]*Beilstein Endowed Chair for Cheminformatics, Johann Wolfgang Goethe-University, Siesmayerstrasse 70, 60323 Frankfurt am Main, Germany*
[2]*Schneider Consulting GbR, George-C.-Marshall Ring 33, 61440 Oberursel, Germany*

Received 9 November 2008; Accepted 22 December 2008
DOI 10.1002/jcc.21218
Published online 5 March 2009 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Measuring the (dis)similarity of molecules is important for many cheminformatics applications like compound ranking, clustering, and property prediction. In this work, we focus on real-valued vector representations of molecules (as opposed to the binary spaces of fingerprints). We demonstrate the influence which the choice of (dis)similarity measure can have on results, and provide recommendations for such choices. We review the mathematical concepts used to measure (dis)similarity in vector spaces, namely norms, metrics, inner products, and, similarity coefficients, as well as the relationships between them, employing (dis)similarity measures commonly used in cheminformatics as examples. We present several phenomena (empty space phenomenon, sphere volume related phenomena, distance concentration) in high-dimensional descriptor spaces which are not encountered in two and three dimensions. These phenomena are theoretically characterized and illustrated on both artificial and real (bioactivity) data.

 J Comput Chem 30: 2285–2296, 2009

Key words: distances; high-dimensional data; chemical descriptors; distance concentration

Introduction

The *similar property principle*,[1] which states that structurally similar molecules tend to have similar properties, constitutes the underlying idea of many applications in cheminformatics such as compound ranking (chemical similarity searching),[2] ligand-based virtual screening,[3] and, diversity analysis for compound library design.[4–6] They are utilized by employing chemical descriptors,[7] i.e., the characterization of molecules by numerical attributes, together with a measure of (dis)similarity defined on the descriptors. The corresponding mathematical abstraction is that of a vector space,[8] resulting in chemical descriptor spaces, often of high dimensionality. Although a successful concept, some problems have been associated with high-dimensional chemical descriptor spaces, e.g., chance correlations in quantitative structure-activity relationships.[9,10] In this work, we focus on phenomena related to the behavior of (dis)similarity measures in high-dimensional real vector spaces and their consequences for similarity-based approaches. Note that binary spaces are vector spaces over the finite field $\mathbb{F}_2 = \{0, 1\}$; since their special structure has been exploited in dedicated work,[11–13] they are not treated here.

Problems related to high-dimensional representations have been recognized both within[9,14] and outside[15,16] of chemistry, e.g., in the database community, where they are relevant to indexing and retrieval. The influence of (dis)similarity measures on cheminformatics tasks has been investigated before.[17] To our knowledge, the present study is unique in its emphasis on high-dimensional chemical descriptor spaces.

In the following, we present several phenomena related to distances in high-dimensional spaces ("Distance Phenomena in High-Dimensional Spaces" section), investigate their impact on cheminformatics applications ("Practical Consequences" section), and, conclude ("Conclusions" section). The mathematical formalism of norms, metrics, and inner products is briefly recapitulated in an appendix, including a table of (dis)similarity measures of relevance to cheminformatics (Table A1).

Dataset

We present empirical results on artificial as well as real data. For the latter, we use a bioactivity dataset, the COBRA database,[18] version 8.6. This commercial dataset contains 10886 small molecules (of which 38 could not be processed by some of the used software and were removed), annotated with activity on 707 biological

Additional Supporting Information may be found in the online version of this article.

Correspondence to: M. Rupp; e-mail: matthias.rupp@bio.uni-frankfurt.de

Table 1. Typical Dataset Sizes n with Maximum Covered Dimension $\max d = \lfloor \log_2(n) \rfloor$.

n	$\max d$	Description
10^2	6	Virtual screening training set
10^4	13	COBRA drug database
10^5	16	Known drugs
10^6	19	High-throughput screening dataset
10^7	23	CAS REGISTRY database[22]

targets and 50 interaction types. Where necessary, e.g., for the investigation of clustering behavior, we divided the dataset into classes based on these annotations. For statistical reasons, we retained only combinations of target and interaction type with at least 30 entries, resulting in 96 classes (e.g., COX-2 inhibitors, PPARγ agonists, etc.; see Supporting Information for a complete list) containing 5066 compounds, 46.7% of the whole dataset.

Two descriptors were used: (a) the CATS2D descriptor,[19,20] normalized by dividing eachcount by the added occurrences of the two respective pharmacophoric types. The CATS2D descriptor assigns pharmacophoric types (hydrogen bond donor (D), hydrogen bond acceptor (A), positive charge (P), negative charge (N), lipophilic (L)) to vertices of the structure graph and then counts shortest path lengths between all pairs of vertices belonging to a pharmacophoric type pair (DD, DA, DP, DN, DL, AA, AP, AN, . . ., LL). Path lengths from 0 up to 9 were considered, resulting in a $15 \cdot 10 = 150$ dimensional vector. Each entry was divided by the added occurences of the two respective pharmacophore types, e.g., if there were 3 DD paths of length 2, and 15 DD paths in total, the entry for DD paths of length 2 would be 0.2. This restricts each component of the vector to the range $[0, 1]$, and no further standardization was applied. On this dataset, 9 out of 150 dimensions were constant, and 9705 samples were unique. (b) all MOE 2D descriptors (Molecular Operating Environment, version 2008.06, Chemical Computing Group, www.chemcomp.com), standardized by subtraction of mean and division by standard deviation. On this dataset, 19 out of 184 dimensions were constant, and 9950 samples were unique.

Distance Phenomena in High-Dimensional Spaces

Norms, metrics, and, inner products (see Appendix) generalize concepts from two- and three-dimensional geometry like length, distance, and angle in a straightforward way to more abstract spaces, in particular to high-dimensional real vector spaces. There, phenomena occur for which low-dimensional geometry provides no intuition. The root cause of these phenomena is that distance is measured across volume, which increases exponentially with dimension.

Empty Space Phenomenon

Consider a finite sample $x_1, \ldots, x_n \in \mathbb{R}^d$. A partitioning of each dimension into two parts, such that each part contains at least one sample, results in a partitioning of $\mathbb{R}^d$ into 2^d compartments. Since the number of compartments grows exponentially with the dimension d, an exponential number of samples is needed to cover $\mathbb{R}^d$ in the sense that each compartment contains at least one sample.

Table 2. Dimensionality d of Commonly Used Descriptor Spaces.

d	Descriptor
$\sim$50	Mini-fingerprints[23,24]
72	VolSurf descriptor[25]
120	Ghose-Crippen fragment descriptors[26]
150	CATS2D pharmacophore descriptor[19,20]
184	MOE (version 2008.06) 2D descriptors

This is a reason why density estimation in high dimensions is difficult.[21] For practical scenarios, almost all of the compartments will be empty. As an example,consider a compound library with 10^8 compounds described by Ghose-Crippen fragment descriptors, for which $d = 120$, a common dimensionality of chemical descriptor spaces (Table 2). Although the dataset is large (Table 1), the fraction of compartments covered is at most $10^8/2^{120} \approx 10^{-28} \approx 0$. The maximum dimension that could be covered by this dataset is $\lfloor \log_2(10^8) \rfloor = 26$. From Tables 1 and 2, it is clear that in typical scenarios, the chemical space spanned by a descriptor will be empty in terms of dataset coverage.

The distribution of the samples is another matter. Compound collections usually exhibit structure due to selection bias, which suggests that they lie on lower-dimensional submanifolds of the descriptor space. Consider n samples drawn independently and uniformly distributed from $[0, 1]^d \subset \mathbb{R}^d$, where each dimension is partitioned into intervals $[0, \frac{1}{2}]$ and $(\frac{1}{2}, 1]$. The probability that at least one compartment is shared by two or more samples is $1 - \binom{m}{n} \frac{n!}{m^n}$, where $m = 2^d$. For the COBRA dataset and the CATS2D descriptor ($n = 9705$, $d = 141$), this is $\approx 1.689 \times 10^{-35} \approx 0$. Rescaling this data to the range $[0, 1]^{141}$, however, leads to 1016 compartments with two samples or more. Values for MOE 2D descriptors ($n = 9950$, $d = 165$) are comparable (1.058×10^{-42}, 1072). This shows that the COBRA compounds were not sampled uniformly and identically distributed from either the CATS2D or MOE 2D descriptor spaces. Lower intrinsic dimensionality of the dataset is also indicated by a principal components analysis (Fig. 1).

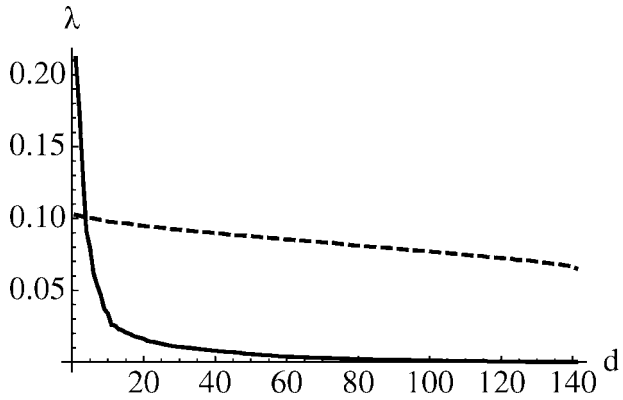

Figure 1. Principal components analysis eigenvalues for the COBRA dataset. Eigenvalues of the empirical covariance matrix for the CATS2D descriptor (solid line), and, 10 886 random samples from $[0, 1]^{141}$ (dotted line) are shown. The sharpest bend of the solid line seems to be at $d = 11$; 90% variance are covered at $d = 50$.

Table 3. Spheres, Balls, and Cubes in d Dimensions.

d	Sphere	Cube
1	$\{-r, +r\}$	Line segment
2	Circle	Square
3	Sphere	Cube
≥4	Hypersphere	Hypercube

A d-dimensional sphere of radius r consists of all points $x \in \mathbb{R}^d$ with $\sum_{i=1}^{d} x_i^2 = r^2$; for a ball, $=$ is replaced by $\leq$. A d-dimensional cube with side length r has points $x \in \mathbb{R}^d$ with $-\frac{r}{2} \leq x_i \leq \frac{r}{2}$ for all i (note that we are not interested in translations and rotations here). Ball and cube can be rewritten as $\left\{x \in \mathbb{R}^d \mid \|x\|_2 \leq r\right\}$ and $\left\{x \in \mathbb{R}^d \mid \|x\|_1 \leq r\right\}$, respectively. The main difference between spheres and cubes therefore lies in the used metric.

Another way to look at this, is that n samples can span (if the embedding space allows it) a subspace of dimension at most $n-1$, but they can only cover a subspace of dimension $\lfloor \log_2(n) \rfloor$. These findings suggest the usefulness of feature selection[27] and dimensionality reduction[28] methods for chemical datasets. Indeed, feature selection is common practice in quantitative structure-activity relationship modeling.

Sphere Volumes

The d-dimensional Euclidean *sphere* $S_{d,r} = \left\{x \mid \|x\|_2 = r\right\}$ of radius $r \geq 0$ (Table 3) has volume[29] $V_{S_{d,r}} = (\pi^{d/2} r^d)/\Gamma(1 + \frac{d}{2})$, where Γ denotes the gamma function.[30] $V_{S_{d,r}}$ goes to zero for $d \to \infty$ (Fig. 2). As a consequence, for a finite sample $x_1, \ldots, x_n \in \mathbb{R}^d$ and fixed radius r, there is a dimension d after which a sphere of radius r centered on x_i contains only x_i and no other sample (Fig. 3). The volume 2^d of a cube circumscribing the unit sphere, on the other hand, goes to infinity. Therefore, a sample drawn uniformly from this cube is almost surely located at its corners (in the cube, but outside of the sphere). For two spheres with different radii $r < r'$, the ratio of their volumes $V_{S_{d,r}}/V_{S_{d,r'}} = \left(\frac{r}{r'}\right)^d$ decreases exponentially with d. Samples drawn uniformly from the larger sphere will therefore lie outside of the smaller sphere with probability $1 - (\frac{r}{r'})^d \xrightarrow{d\to\infty} 1$. This demonstrates the influence of metric and dimension when computing with spherical neighborhoods.

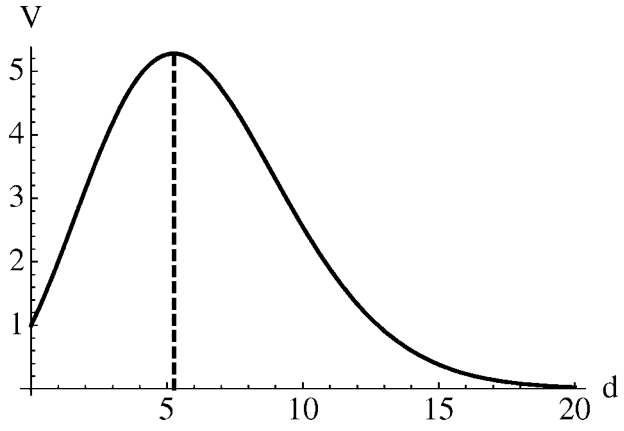

Figure 2. Volume $V_{S_{d,1}}$ of the unit sphere as a function of the dimension d. The maximum is at ≈5.257, and, $V_{S_{5,1}} = \frac{8}{15}\pi$.

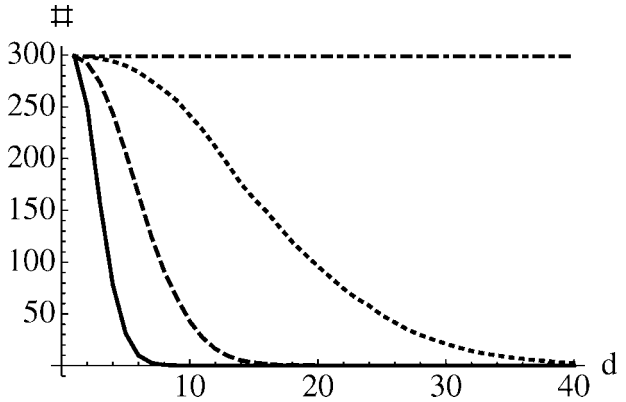

Figure 3. Average number of samples included in a unit sphere centered on each of 300 random samples in $[0,1]^d$. Shown are L^p-norm based metrics for $p = 1$ (solid line), $p = 2$ (dashed line), $p = 4$ (dotted line); in the limit $p \to \infty$ (dash-dotted line), the unit sphere becomes a cube.

Phenomena related to spherical or ellipsoidal volumes also affect statistics. Consider a d-dimensional standard normal distribution, i.e., a distribution which generates samples $x_i \in \mathbb{R}^d$ with independent components $(x_i)_j \sim \mathcal{N}(0,1)$. With increasing dimension, the probability mass contained in a sphere of fixed radius around the origin decreases rapidly (Fig. 4). While in one dimension most points lie close to the origin, in higher dimensions almost no point does; in this sense, for high dimensions most of the probability mass lies in the tails and not in the center of a normal distribution. This behavior is due to L^p-norms being defined in terms of absolute component values $|(x_i)_j|$, with $\mathrm{E}\left(|(x_i)_j|\right) = \sqrt{2/\pi}$ causing $\|x_i\|_p$ to grow with each added dimension. Since the distribution of $\|x_i\|_p$ is unimodal, the samples tend to lie on a hypersphere with radius $r = \mathrm{E}\left(\|x_i\|_p\right)$. The value of r depends on d and p:

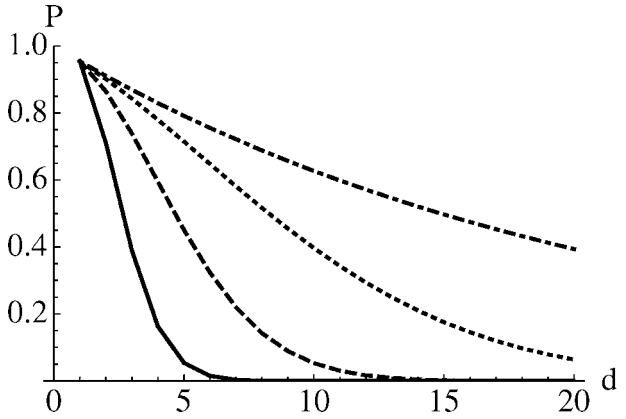

Figure 4. Probability mass of the d-dimensional standard normal distribution contained in a sphere of radius 2, as measured by the grid norm (solid line), the Euclidean norm (dashed line), the L^4-norm (dotted line), and, the max-norm (dash-dotted line). Numerical estimation with 10^6 samples.

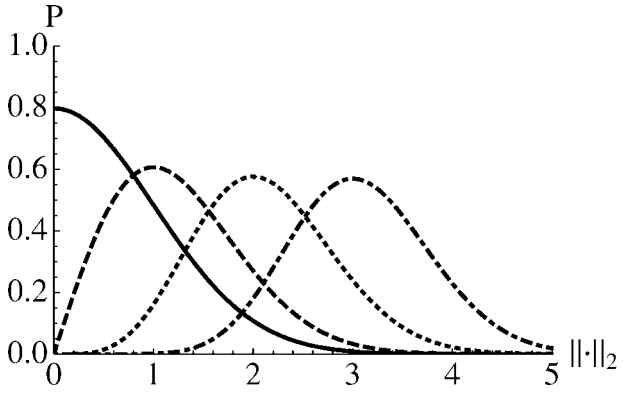

Figure 5. Distribution of the Euclidean norm of samples drawn from a standard normal distribution in d dimensions. The L^2-norm follows a χ-distribution with d degrees of freedom. Shown are probability densities for $d = 1$ (solid line), $d = 2$ (dashed line), $d = 5$ (dotted line), and, $d = 10$ (dash-dotted line).

for the grid norm, $\mathrm{E}(\|x_i\|_1) = d\sqrt{2/\pi}$; for the Euclidean norm, $\mathrm{E}(\|x_i\|_2) = \sqrt{2}\Gamma(\frac{d+1}{2})/\Gamma(\frac{d}{2})$ (Fig. 5).

Distance Concentration

The concentration of norms is not limited to normally distributed samples, and it also affects the distances between samples. In high-dimensional spaces, under mild assumptions, sample norms tend to concentrate. As a consequence, all distances are similar, samples lie on a hypersphere, and, each sample is nearest neighbor of all other samples. For an intuitive explanation, consider independent samples $x_1, \ldots, x_n$ drawn uniformly from $[0, 1] \subset \mathbb{R}$. For $n \to \infty$, $\mathrm{E}(x_i) = 0.5$ because the values average out over the samples. Now consider a single sample $x \in [0, 1]^d$ for $d \to \infty$. Again, the values average out, but this time over the components of x, so $\lim_{d\to\infty} \frac{\|x\|_1}{d} = 0.5$. Note that L^p-norms increase with d. Figure 6 shows this for different norms on artificial data.

The concentration of L^p-norms and associated Minkowski metrics has been formally studied,[16,31–33] often using the (*absolute*) *contrast* $\max_i \|x_i\| - \min_i \|x_i\|$ and the *relative contrast* $\frac{\max_i \|x_i\| - \min_i \|x_i\|}{\min_i \|x_i\|}$ as measures of concentration. However, these depend on extremal values, and therefore on sample size, and are highly volatile. Instead, we use another measure of spread versus location, the *variation coefficient* (or *relative variance*[16]) $\frac{\sigma}{\mu} = \frac{\sqrt{\mathrm{var}(\|\cdot,\cdot\|)}}{\mathrm{E}(\|\cdot,\cdot\|)}$, where small values indicate concentration. Note that $\frac{\sigma}{\mu}$ can equivalently be defined in terms of norms by a change of domain.[16]

Let $X \in \mathbb{R}^d$ be a random variable with independent and identically distributed components. Then,[16]

$$\lim_{d\to\infty} \frac{\mathrm{E}(\|X\|_p)}{d^{\frac{1}{p}}} = c, \quad \lim_{d\to\infty} \frac{\mathrm{var}(\|X\|_p)}{d^{\frac{2}{p}-1}} = c', \quad \text{and,}$$

$$\lim_{d\to\infty} \frac{\sqrt{\mathrm{var}(\|X\|_p)}}{\mathrm{E}(\|X\|_p)} = 0, \quad (1)$$

where c and c' are constants not depending on d. This shows that, under the strong assumption of independence and identical distribution of the components, all L^p-norms and Minkowski distances concentrate and also gives the rates of growth $\mathrm{E}(\|X\|_p) \sim c\, d^{1/p}$ and $\mathrm{var}(\|X\|_p) \sim c' d^{2/p-1}$ (Fig. 6). Note that c and c' depend on p. Equation (1) stays valid[16] for differently distributed components and dependences between them. In the first case, the equation holds if the data are standardized, i.e., have zero mean and unit variance (subtracting the mean and dividing by the standard deviation achieves this). Standardization ensures that the norm is not dominated by a few components. In the second case, concentration takes place but depends on the intrinsic dimensionality of the data, as opposed to the dimensionality of the vector space itself.

Chemical descriptor spaces are, as a rule, normalized in some form or other, and from Table 2 and Figures 2–4 and 6, it is clear that their dimensionality is high enough for distance concentration to occur. However, due to dependencies between descriptors and due to selection bias (see discussion of empty space phenomenon), the intrinsic dimensionality of chemical datasets will be lower than the dimensionality of the embedding descriptor space. Table 4 lists variation coefficients using different (dis)similarity measures on the COBRA dataset.

Fractional distances, i.e., Minkowski metrics with $0 < p < 1$ (these do not satisfy the triangle inequality anymore), have been proposed based on the fact that for independent and uniformly distributed data, (relative) contrast and variation coefficient improve with decreasing p.[33] However, this is not the case for other distributions.[16] Figure 7 shows an example for which σ/μ increases

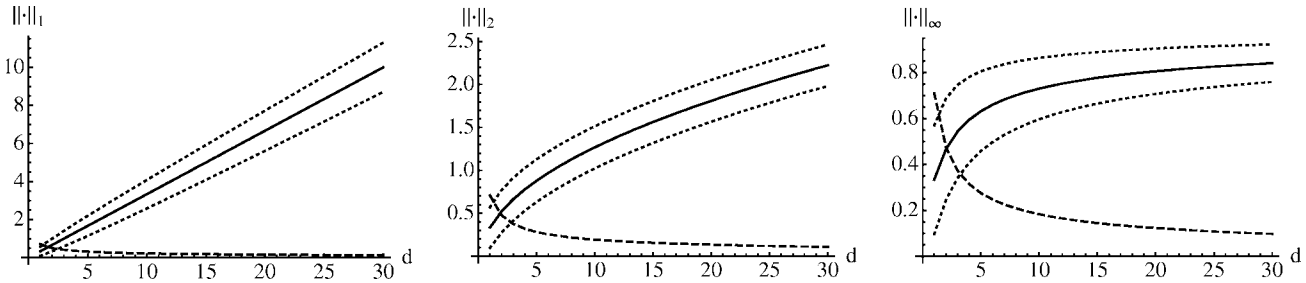

Figure 6. Behavior of L^p-norm based distances. For $n = 10^5$ distances between points sampled uniformly and independently from $[0, 1]^d$, the mean (solid line) ± its standard deviation (dotted lines), and, the coefficient of variation (dashed line) are shown for the Manhattan distance (left), the Euclidean distance (middle), and the maximum metric (right).

Table 4. Variation Coefficients for Different (dis)Similarity Measures on COBRA Dataset.

	Fractional distances					Minkowski metrics					Similarity coefficients			
Descriptor	1/10	1/4	1/2	3/4	9/10	1	2	3	5	∞	$1-r$	$1-d$	$1-t$	$1-c$
CATS2D	3.32	1.01	0.51	0.36	0.32	0.30	0.21	0.19	0.18	0.21	0.34	0.35	0.25	0.36
MOE 2D	0.70	0.52	0.46	0.43	0.42	0.42	0.38	0.37	0.40	0.49	0.29	0.30	0.17	0.33

r = Pearsons correlation, d = Dice coefficient, t = Tanimoto coefficient, c = Carbó index.

with p. For chemical datasets, which are unlikely to be independent and uniformly distributed, fractional distances do not necessarily result in a higher variation coefficient.

Practical Consequences

Virtual screening is often performed in high-dimensional descriptor spaces. It is not clear how exactly distance concentration affects the performance of distance-based algorithms like, e.g., k-nearest neighbor clustering, ranking of a database against a reference compound, or, dimensionality reduction methods trying to preserve local distances. Here, we investigated the influence of dimensionality and (dis)similarity measure on some practical cheminformatics tasks.

Compound Ranking

We investigated the influence of (dis)similarity measures on the ranking of chemical compounds against a reference compound. To compare two (dis)similarity measures, each compound in the 96 selected COBRA classes was taken as reference compound and the other compounds of the class were ranked against it. Spearmans rank correlation coefficient[34] was computed for the generated pairs of lists. Table 5 presents the results using MOE 2D descriptors; results for the CATS2D descriptor were similar.

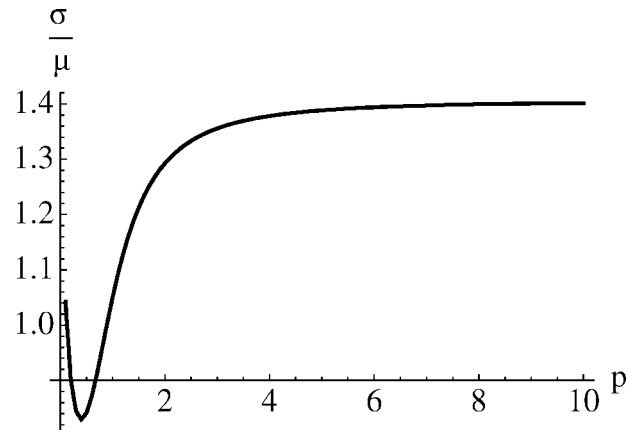

Figure 7. Example of variation coefficient increasing with L^p-norm parameter p. The variation coefficient of the L^p-norm of 10^6 samples in $\mathbb{R}^{20}$ as a function of p is shown. Each sample component is drawn independently from the 5th power of a standard normal distribution, $(x_i)_j = X^5$ where $X \sim \mathcal{N}(0,1)$. Raising the normal distribution to a higher power results in a larger range of extreme values; this favors norms whose value depends more on larger vector components.

The Minkowski metrics and the similarity coefficients exhibit strong correlation among themselves, but weaker correlation between them. Within the former, correlation declines and variance increases with increasing difference in p. Within the latter, correlation is constantly high, with low variance. We attribute the difference in correlation between the two groups to a basic difference in their concepts: The first group measures feature differences ($|x_i - y_i|$, see Table A1), whereas the other group measures the presence of (common) features ($x_i \cdot y_i$). To see that the two groups behave differently, consider $x = (a, \ldots, a)$, $y = (b, \ldots, b) \in \mathbb{R}^d$. The Manhattan distance of x and y is $d|a - b|$; the similarity coefficients are scaled versions of the inner product $\langle x, y\rangle = dab$. For $a = b \gg 1$, the former is zero, while the latter is large; for $da > 0$, $b = \frac{a}{a+1}$, both equal $d\frac{a^2}{a+1} \approx a$. Figures 8–11 present examples. It has been noted before[17] that by using more than one (dis)similarity measure for compound ranking one can retrieve a more diverse set of actives.

Clustering

With increasing dimension, samples become equidistant to each other due to distance concentration. Beyer et al.[31] argue that if the samples are clustered, it is possible to at least discriminate between inter-cluster and intra-cluster distances, as long as the query sample is close to a single cluster. However, it is not possible to choose meaningfully within a cluster, nor to choose a cluster in a meaningful way if the query point is not close to a single cluster.

Table 6 lists average distances within (intra) and between (inter) classes for the COBRA dataset, averaged over classes, using both descriptors and various (dis)similarity measures. In no case were intra- and inter-class distances separable in the sense that their means were further apart than the sum of their respective standard deviations. In 16 out of 19 cases, the inter-class mean was within the mean plus standard deviation of the intra-class distances. For other class definitions (by target only and by a very rough division into enzymes, G protein-coupled receptors, ion channels, kinases, nuclear receptors, proteases, and proteins) results were similar. Figure 12 shows corresponding histograms. These results indicate that classification based on nearest-neighbor distances of the COBRA dataset using high-dimensional descriptor spaces is problematic due to the close proximity of mean inter- and intra-class distances. Indeed, experiments with 1-NN classification yielded about 60% correct classifications and 30% wrong classifications; in 10% of the cases, the query sample was equidistant to two clusters. Results for fractional norms were not better than those of the other (dis)similarity measures. From the investigated (dis)similarity measures, the Carbó index yielded the best separation, the difference being minor.

Table 5. Average Spearman Rank Correlation Coefficient (cc) Between (dis)Similarity Measures on COBRA Dataset with MOE 2D Descriptors.

	Fractional and Minkowski distances p						Similarity coefficients		
cc	$\frac{1}{10}$	$\frac{1}{2}$	1	2	5	∞	Pearson	Tanimoto	Carbó
$\|\cdot,\cdot\|_{1/10}$	1	0.90 ± 0.07	0.85 ± 0.09	0.77 ± 0.13	0.54 ± 0.21	0.32 ± 0.24	0.56 ± 0.25	0.60 ± 0.25	0.61 ± 0.25
$\|\cdot,\cdot\|_{1/2}$		1	0.98 ± 0.01	0.90 ± 0.07	0.64 ± 0.19	0.38 ± 0.25	0.61 ± 0.26	0.66 ± 0.28	0.67 ± 0.27
$\|\cdot,\cdot\|_1$			1	0.95 ± 0.04	0.71 ± 0.17	0.43 ± 0.24	0.63 ± 0.26	0.68 ± 0.27	0.69 ± 0.26
$\|\cdot,\cdot\|_2$				1	0.84 ± 0.11	0.54 ± 0.22	0.64 ± 0.25	0.67 ± 0.26	0.68 ± 0.25
$\|\cdot,\cdot\|_5$					1	0.81 ± 0.14	0.51 ± 0.25	0.51 ± 0.25	0.53 ± 0.24
$\|\cdot,\cdot\|_\infty$						1	0.30 ± 0.26	0.30 ± 0.27	0.31 ± 0.26
Pearson							1	0.95 ± 0.05	0.96 ± 0.05
Tanimoto								1	0.98 ± 0.04
Carbó									1

For each pair of (dis)similarity measures, all 5066 compounds of the 96 retained COBRA classes were ranked against all other members of their respective class. Similarity coefficients c were transformed using $1 - c$. Shown are mean $\pm$ standard deviation. The Dice coefficient obtained results identical to those of the Tanimoto coefficient.

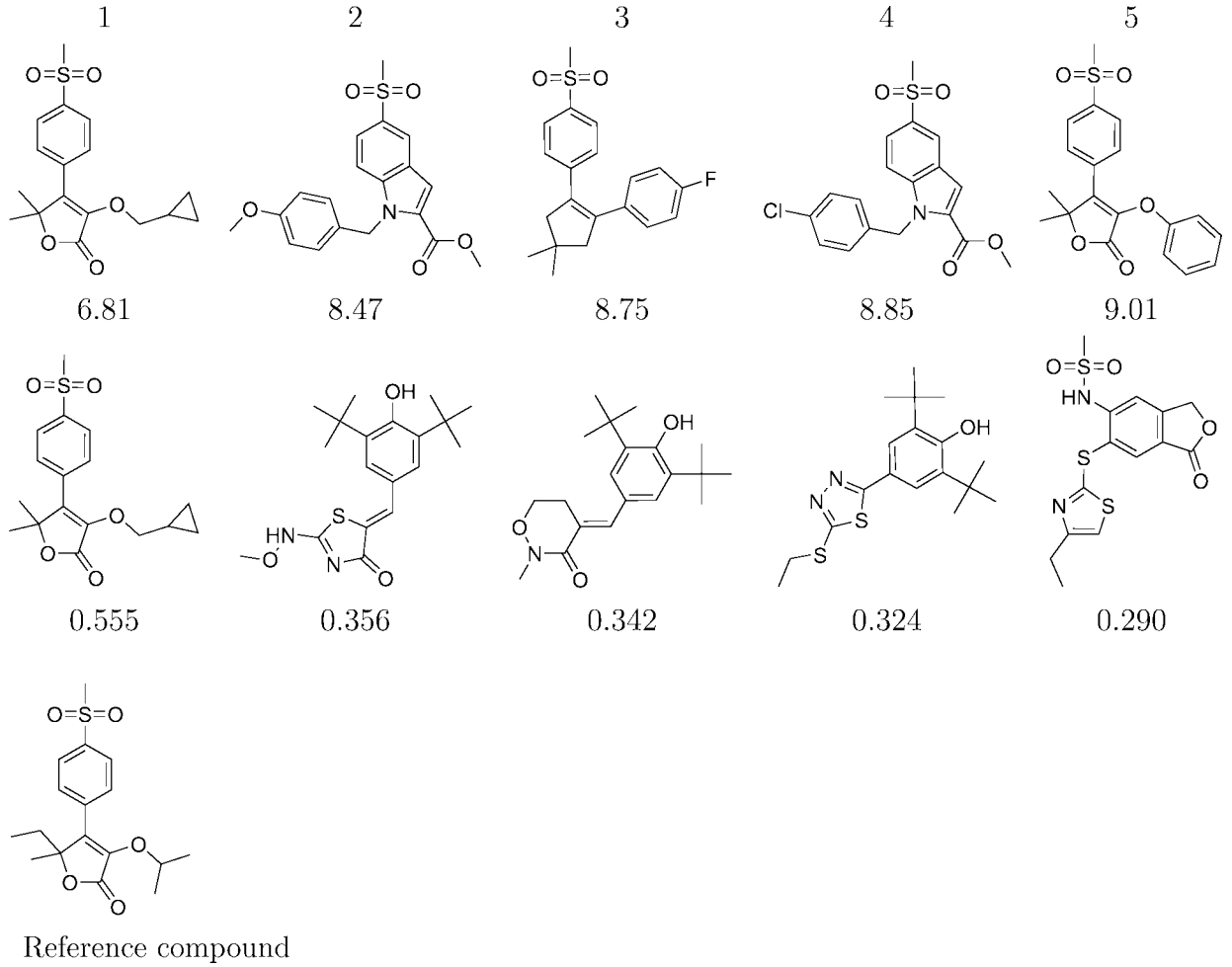

Figure 8. Compound ranking of COBRA (MOE 2D) selective COX-2 inhibitors against a reference compound. The five most similar compounds as measured by the Euclidean distance (upper row) and the Tanimoto similarity coefficient (lower row) are shown. Only the first compound is selected by both (dis)similarity measures.

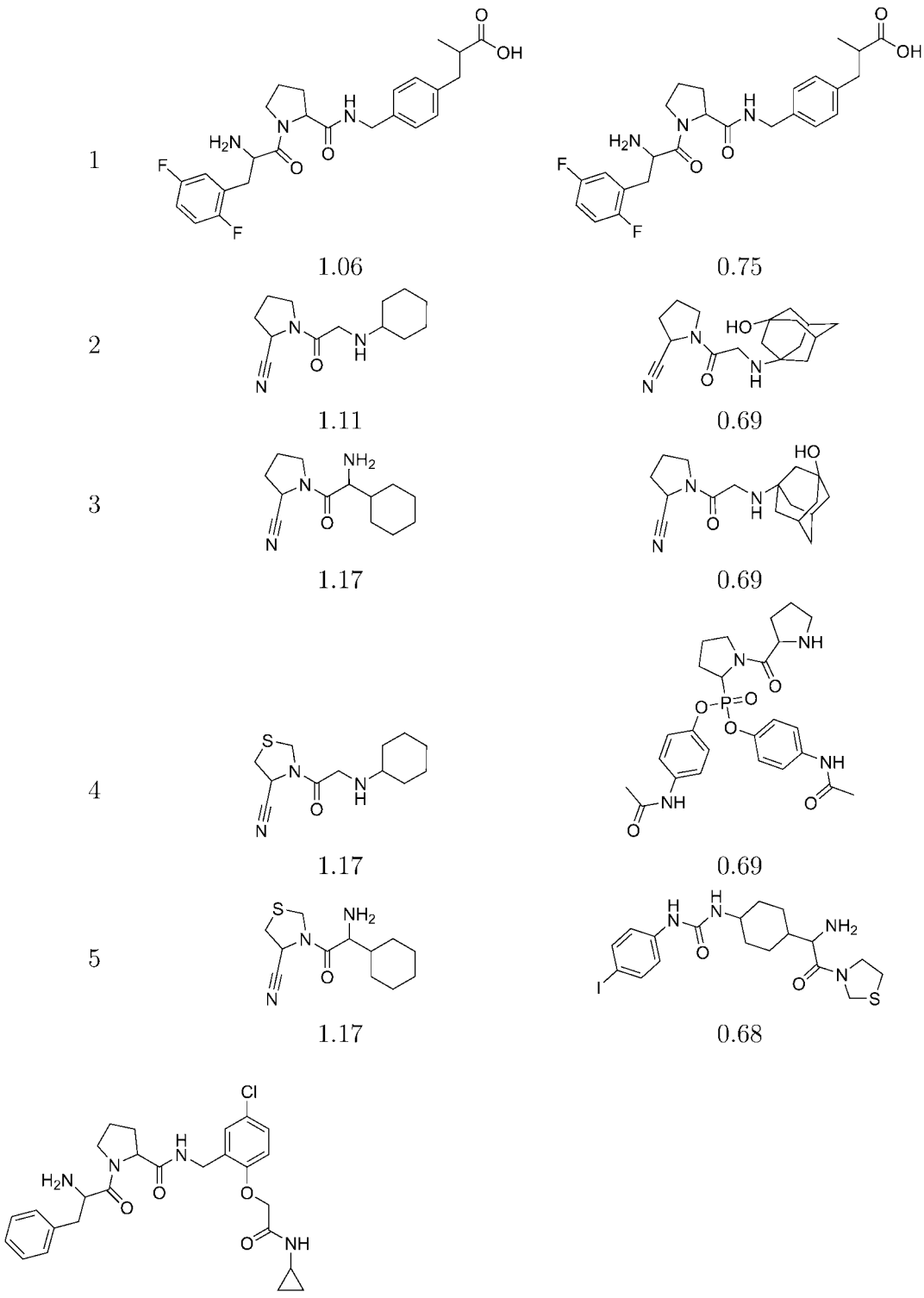

Figure 9. Compound ranking of COBRA (CATS2D) dipeptidyl peptidase IV (DPP-IV) inhibitors. The five most similar compounds as measured by the Euclidean distance (left column) and the Tanimoto similarity coefficient (right column), as well as the reference compound are shown. Only the first compound is selected by both (dis)similarity measures.

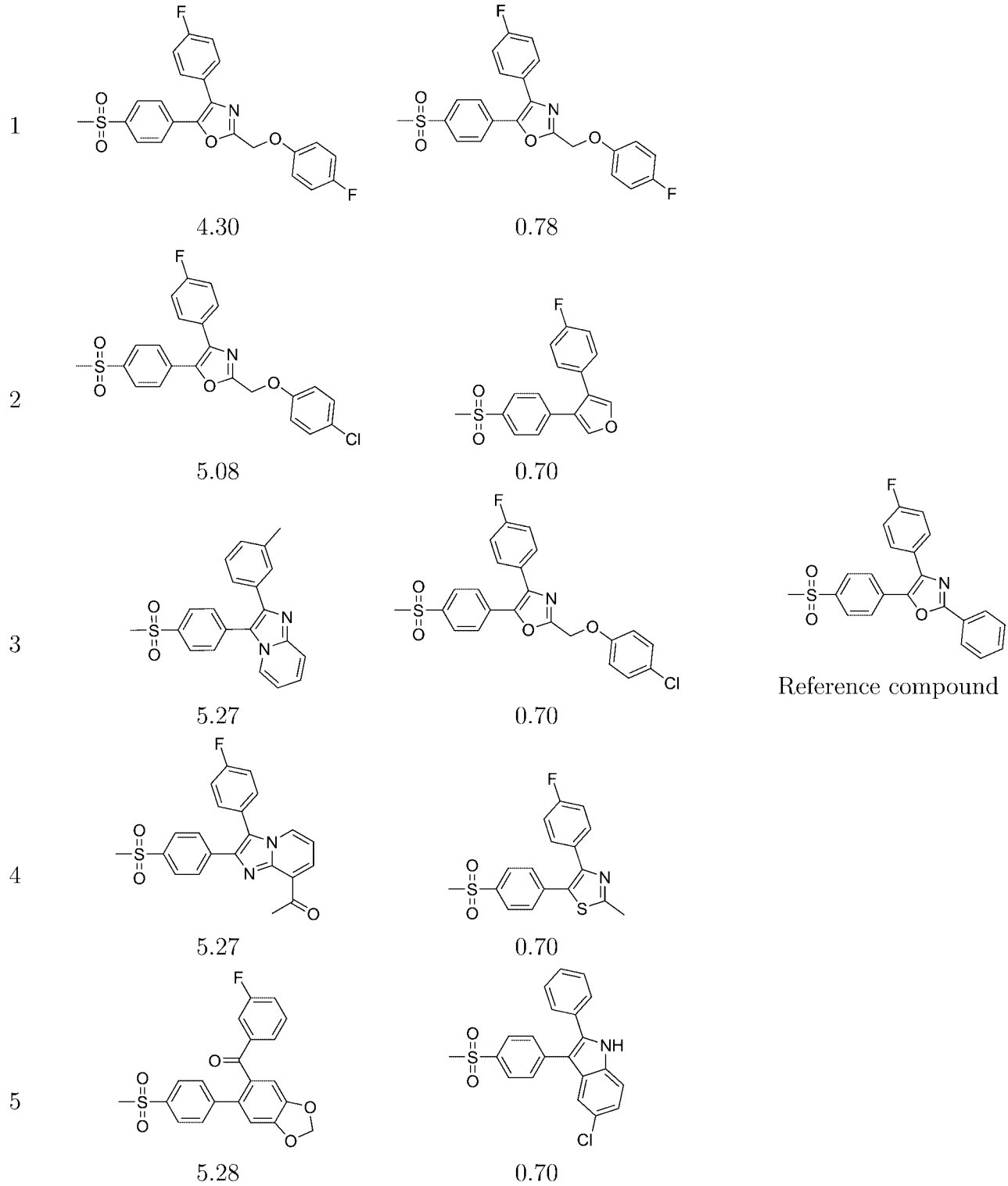

Figure 10. Compound ranking of COBRA (MOE 2D) cyclooxygenase-2 (COX-2) selective inhibitors. The five most similar compounds as measured by the Euclidean distance (left column) and the Tanimoto similarity coefficient (right column), as well as the reference compound are shown. Two compounds were selected by both (dis)similarity measures.

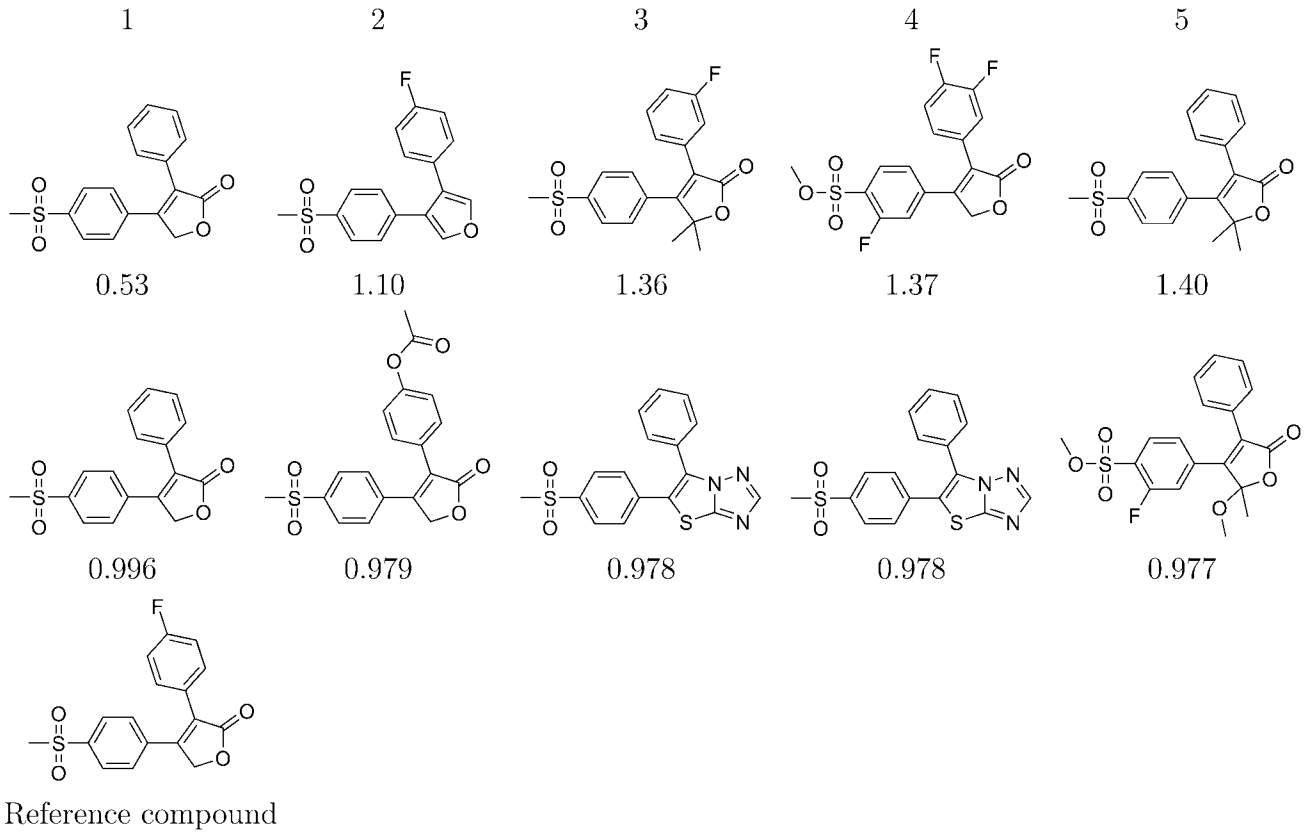

Figure 11. Compound ranking of COBRA (CATS2D) cyclooxygenase-2 (COX-2) selective inhibitors. The five most similar compounds as measured by the Manhattan distance (top row) and the Carbó similarity coefficient (bottom row), as well as the reference compound are shown. Two compounds were selected by both (dis)similarity measures.

Conclusions

We described several phenomena of distances in high-dimensional spaces, namely, the empty space phenomenon, sphere volume related phenomena, and, distance concentration, all of which have the exponential growth of volume with dimension as their root cause. Consequences for compound ranking and clustering were investigated on real and artificial data.

The described effects set in very early, at dimensions between 5 and 20. From the dimensionality of common chemical descriptor spaces, it is clear that the occurrence of these phenomena is rule rather than exception. Since our geometrical intuitions are based

Table 6. Intra- and Inter-Class Distances (mean ± standard deviation, averaged over classes) in COBRA Dataset.

	CATS2D		MOE 2D	
Measure	Intra	Inter	Intra	Inter
$\|\cdot,\cdot\|_{3/4}$	27.27 ± 10.53	35.75 ± 11.77	581.22 ± 261.95	798.04 ± 318.40
$\|\cdot,\cdot\|_1$	7.95 ± 2.66	10.10 ± 2.71	120.00 ± 51.20	162.00 ± 61.90
$\|\cdot,\cdot\|_2$	1.42 ± 0.37	1.71 ± 0.32	13.00 ± 4.74	16.60 ± 5.62
$\|\cdot,\cdot\|_3$	0.87 ± 0.21	1.02 ± 0.17	6.94 ± 2.38	8.52 ± 2.70
$\|\cdot,\cdot\|_4$	0.71 ± 0.17	0.82 ± 0.13	5.39 ± 1.85	6.42 ± 2.03
$\|\cdot,\cdot\|_\infty$	0.52 ± 0.13	0.58 ± 0.11	3.88 ± 1.53	4.34 ± 1.66
Pearson	0.33 ± 0.14	0.45 ± 0.14	0.66 ± 0.29	1.00 ± 0.28
Dice	0.27 ± 0.12	0.38 ± 0.12	0.66 ± 0.30	0.99 ± 0.30
Tanimoto	0.41 ± 0.14	0.53 ± 0.12	0.75 ± 0.24	0.97 ± 0.17
Carbó	0.26 ± 0.11	0.36 ± 0.11	0.65 ± 0.31	1.00 ± 0.32

Similarity coefficients c were converted to dissimilarity measures using $1 - c$. Results for broader class definitions were similar.

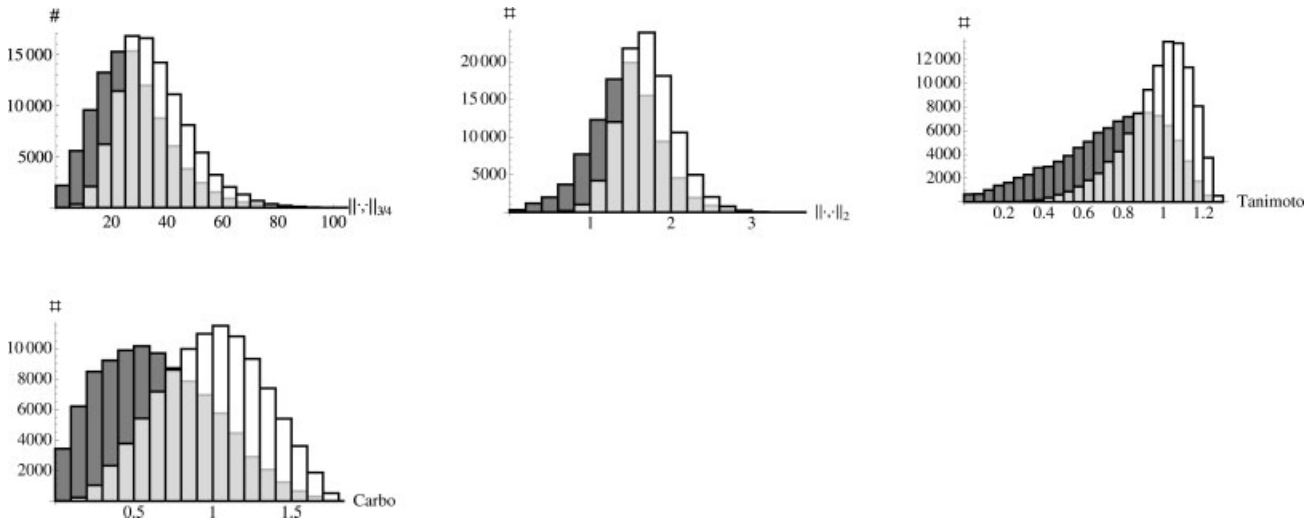

Figure 12. Histograms of intra- and inter-class distances on COBRA dataset. Histograms for intra-class (dark color) and inter-class (light color) distances, each with 10^5 samples, using CATS2D with fractional distance $\|\cdot,\cdot\|_{3/4}$ (top left), CATS2D with Euclidean distance (top middle), MOE 2D with $1 -$ Tanimoto coefficient (top right), and, MOE 2D with $1 -$ Carbó index (bottom left) are shown. Results for broader class definitions were similar. Previous results[35] on an earlier version of the COBRA dataset showed higher separation.

on two- and three-dimensional spaces, care needs to be taken when computing in spaces of high dimensionality.

We demonstrated the influence of the choice of (dis)similarity measure, with major differences between Minkowski metrics and similarity coefficients. For compound ranking, we recommend to use more than one (dis)similarity measure, preferably at least one Minkowski metric and one similarity coefficient, since this increases the diversity of the resulting structures. The parameter p of Minkowski metrics (or fractional distances) can be chosen by plotting the variation coefficient as a function of p. For labeled samples, p can be chosen by a maximal separation criterion.

For uniformly distributed data, the variation coefficient decreases with p. While fractional distances improve the variation coefficient in some cases, they do not satisfy the triangle inequality, which may require adjustment of algorithms. If this is not a concern, fractional distances are a valid choice. In the face of norm concentration (all samples lying on a hypersphere), the Carbó index, which corresponds to the distance between samples projected onto the unit sphere (spherical distance), seems a reasonable choice; in accordance, the Carbó index showed better separation capability between intra- and inter-class distances than the other similarity coefficients.

We further argued that chemical datasets are likely to occupy lower-dimensional manifolds in chemical descriptor spaces. This suggests the usefulness of feature selection and dimensionality reduction methods.

In this work, we have not addressed more advanced concepts which have not entered the field of cheminformatics yet. In distance metric learning,[36] for example, the metric is adapted to the dataset. One metric used for this purpose is the Mahalanobis metric (Table A1). By solving a convex optimization problem, the weighting matrix M is chosen such that distances between samples with different labels (inter-class distances) are larger than distances between samples with the same label (intra-class distances).[37] This approach was demonstrated to be highly competitive in k-nearest neighbor classification (e.g., large margin nearest neighbors[38]). Many advanced machine learning techniques are available, and the virtual screening community would benefit from such a method transfer.

Acknowledgments

The authors thank Andreas Schüller and Yusuf Tanrikulu for helpful discussion, as well as the anonymous referees for their constructive comments. This research was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften and a scholarship for M. R. from the Frankfurt International Research Graduate School for Translational Biomedicine (FIRST).

Appendix: Measures of (Dis)similarity

We briefly recapitulate formal notions related to the measurement of (dis)similarity, i.e., norms, metrics, inner products, and, similarity coefficients. Table A1 lists examples of relevance to cheminformatics.

A *norm* measures the length, size, or extent of an object. For a vector space $\mathcal{X}$ over a field $\mathbb{K}$, a function $\|\cdot\| : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a norm iff for all $x, y \in \mathcal{X}$ and $\alpha \in \mathbb{K}$ holds

- $\|x\| \geq 0$ (non-negativity) and $\|x\| = 0 \Leftrightarrow x = 0$
- $\|\alpha x\| = |\alpha|\|x\|$ (homogeneity)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality, subadditivity)

Figure A1 shows the unit sphere (see "Sphere Volumes" section) as measured by L^p norms (Table A1) for different values of p. If the condition $\|x\| = 0 \Leftrightarrow x = 0$ is dropped, $\|\cdot\|$ is a *seminorm*. An example of a seminorm is $\|f\| = |f(0)|$ for $\mathbb{K} = \mathcal{C}^0$, the space of continuous functions. On finite dimensional vector spaces, all norms are equivalent in the sense that $\exists \gamma_1, \gamma_2 > 0 : \forall x \in \mathcal{X} : \gamma_1 \|x\|_a \leq \|x\|_b \leq \gamma_2 \|x\|_a$, or, equivalently, $\gamma_1 \leq \|x\|_a / \|x\|_b \leq \gamma_2$. *Matrix norms* satisfy the additional requirement

Table A1. Common Norms, Metrics, Inner Products, and Similarity Coefficients.

Norm $\|x\|$, $\|A\|$	Domain	Name
$(\sum_{i=1}^{m} \|x_i\|^p)^{1/p}$, $p \geq 1$	$\mathbb{K}^m$	L^p-norm $\|\cdot\|_p$
$\sum_{i=1}^{m} \|x_i\|$	$\mathbb{K}^m$	L^1-norm $\|\cdot\|_1$, grid norm, sum norm
$\sqrt{\sum_{i=1}^{m} \|x_i\|^2}$	$\mathbb{K}^m$	L^2-norm $\|\cdot\|_2$, Euclidean norm
$\max_{1\leq i\leq m} \|x_i\|$	$\mathbb{K}^m$	L^∞-norm $\|\cdot\|_\infty$, max norm
$\sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m} \|a_{i,j}\|^2} = \sqrt{\mathrm{tr}(AA^T)}$	$\mathbb{R}^{k\times m}$	Frobenius norm
$\max_{\|x\|=1} \|Ax\|$	$\mathbb{R}^{k\times m}$	Matrix norm induced by $\|\cdot\|$
Metric $m(x, y)$	**Domain**	**Name**
$(\sum_{i=1}^{m} \|x_i - y_i\|^p)^{1/p}$, $p \geq 1$	$\mathbb{K}^m$	L^p norm induced metric, Minkowski distance
$\sum_{i=1}^{m} \|x_i - y_i\|$	$\mathbb{K}^m$	Manhattan metric (L^1-norm based)
$\sqrt{\sum_{i=1}^{m} \|x_i - y_i\|^2}$	$\mathbb{K}^m$	Euclidean metric (L^2-norm based)
$\max_{1\leq i\leq m} \|x_i - y_i\|$	$\mathbb{K}^m$	Maximum (Chebyshev) metric (L^∞-norm based)
$\sqrt{(x-y)^T M(x-y)}$, M s. p. d.	$\mathbb{K}^m$	Mahalanobis metric
Inner product $\langle x, y\rangle$, $\langle f, g\rangle$	**Domain**	**Name**
$x^T y = \sum_{i=1}^{m} x_i y_i$	$\mathbb{R}^m$	Standard inner product, dot product
$\mathrm{tr}(A^T B)$	$\mathbb{R}^{k\times m}$	Matrix standard inner product
$x^T My$, M s. p. d.	$\mathbb{R}^m$	Weighted inner product
$\int_a^b f(t)g(t)\mathrm{d}t$	$C[a, b]$	Inner product of continuous functions on an interval
$\int_{\mathcal{X}} f(t)g(t)\mathrm{d}t$	L^2	Inner product of square integrable functions
Similarity coefficient	**Range**	**Name**
$\frac{\mathrm{covar}\,(x,y)}{\sqrt{\mathrm{var}\,(x)\,\mathrm{var}\,(y)}}$	$[-1, 1]$	Product-moment (Pearsons) correlation coefficient
$\frac{2\,\langle x,y\rangle}{\langle x,x\rangle + \langle y,y\rangle}$	$[-1, 1]$	Hodgkin index, Dice coefficient
$\frac{\langle x,y\rangle}{\langle x,x\rangle + \langle y,y\rangle - \langle x,y\rangle}$	$\left[-\frac{1}{3}, 1\right]$	Tanimoto coefficient
$\frac{\langle x,y\rangle}{\sqrt{\langle x,x\rangle\,\langle y,y\rangle}} = \frac{\langle x,y\rangle}{\|x\|_2\,\|y\|_2}$	$[-1, 1]$	Carbó index, cosine similarity

For Minkowski distances with $0 \leq p < 1$, the triangle inequality is reversed. The Frobenius matrix norm is the L^2-norm applied to the concatenated rows or columns of a matrix A. All similarity coefficients are over the domain $\mathbb{R}^m$. A^T = transpose of matrix A, $\mathrm{tr}(A)$ = trace of matrix A, $\mathrm{covar}\,(x, y)$ = covariance of random variables x and y, s.p.d. = symmetric positive definite.

- $\|AB\| \leq \|A\|\|B\|$ (submultiplicativity)

for all matrices A, B of compatible shape, including the case of A or B being a vector.

A *metric* measures the distance between or dissimilarity of two objects. A function $m : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined over some set $\mathcal{X}$ is a metric iff for all $x, y \in \mathcal{X}$

- $m(x, y) \geq 0$ (non-negativity) and $m(x, y) = 0 \Leftrightarrow x = y$
- $m(x, y) = m(y, x)$ (symmetry)
- $m(x, y) + m(y, z) \geq m(x, z)$ (triangle inequality)

The pair $(\mathcal{X}, m)$ is called a *metric space*. If the condition $m(x, y) = 0 \Leftrightarrow x = y$ is dropped, $m(\cdot, \cdot)$ is a *pseudo-metric*. An example of a pseudo-metric is the *cosine distance* $\arccos(\langle x, y\rangle / \sqrt{\langle x, x\rangle\,\langle y, y\rangle})$, which measures the angle between two vectors in radians. A metric can be constructed from a norm by $m(x, y) = \|x - y\|$; vice versa, a norm can be constructed from a metric via $\|x\| = m(x, 0)$. Norms and metrics constructed in this way obey for all $x, y, z \in \mathcal{X}$, $\alpha \in \mathbb{K}$,

$$m(x + z, y + z) = m(x, y) \quad \text{and} \quad m(\alpha x, \alpha y) = |\alpha| m(x, y).$$

The *inner product* generalizes geometric concepts like length, angle, and, orthogonality. For a real vector space $\mathcal{X}$, a function $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is an inner product iff for all $x, y, z \in \mathcal{X}$, $\alpha \in \mathbb{R}$ holds

- $\langle x, x\rangle \geq 0$ (non-negativity) and $\langle x, x\rangle = 0 \Leftrightarrow x = 0$
- $\langle x, y\rangle = \langle y, x\rangle$ (symmetry)
- $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$ and $\langle \alpha x, y\rangle = \alpha\,\langle x, y\rangle$ (linearity)

The pair $(\mathcal{X}, \langle \cdot, \cdot \rangle)$ is called an *inner product space*. Two vectors $x, y \in \mathcal{X}$ are *orthogonal* iff their inner product is zero, $x \perp y \Leftrightarrow \langle x, y\rangle = 0$. In a real inner product space $\mathcal{X}$, the *angle* (measured in radians) between two non-zero $x, y \in \mathcal{X}$ is defined as

$$\theta \in [0, \pi] \quad \text{such that} \quad \cos\theta = \frac{\langle x, y\rangle}{\|x\|\,\|y\|}.$$

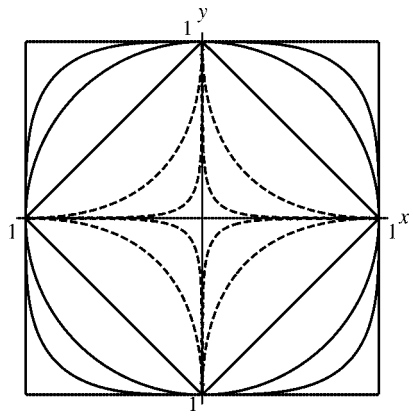

Figure A1. The two-dimensional unit sphere measured by the L^p-norm for $p \in \{\frac{2}{7}, \frac{1}{2}\}$ (dashed lines, inside to outside) and $p \in \{1, 2, \frac{7}{2}, \infty\}$ (solid lines, inside to outside). Note that for $p < 1$, the triangle inequality is reversed.

An inner product $\langle \cdot, \cdot \rangle$ can be used to construct a norm (and corresponding metric) via $\|x\| = \sqrt{\langle x, x \rangle}$. Conversely, given a norm $\|\cdot\|$ on $\mathcal{X}$, $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$ iff the *parallelogram identity* $\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2)$ holds; the Euclidean norm is the only L^p-norm satisfying this identity.[8]

There are useful measures of (dis)similarity which do not fit into the above categories. *Similarity coefficients* often have values in the range $[-1, 1]$, with -1 indicating a strong but adverse relationship, 0 indicating no relationship and 1 indicating a strong relationship (similarity). If $c \in [-1, 1]$, then $1-c \in [0, 2]$, with 0 indicating maximum similarity and 2 indicating maximum dissimilarity. Similarity coefficients often obey some of the metric or other useful properties. For example, the *Kullback-Leibler divergence* (or *relative entropy*) $\sum_{i=1}^{d} p_i \ln(p_i/q_i)$, where $p, q \in \mathbb{R}^d_{\geq 0}$, $\sum_i p_i = \sum_i q_i = 1$, are discrete probability distributions, is not symmetric, but non-negative and equals zero only if $p = q$; it is also convex. Other similarity measures like *Tversky similarity*[39] are defined in set theoretic terms.

References

1. Johnson, M.; Maggiora, G., Eds. Concepts and Applications of Molecular Similarity; Wiley: New York, 1990.
2. Willett, P. J Chem Inform Comput Sci 1998, 38, 983.
3. Böhm, H.-J.; Schneider, G. Virtual Screening for Bioactive Molecules: Wiley-VCH, New York, USA, 2000.
4. Willett, P. Curr Opin Biotechnol 2000, 11, 85.
5. Clark, R. In Combinatorial Library Design and Evaluation; Ghose, A.; Viswanadhan, New York, USA, V., Eds.; CRC Press: 2001 pp. 337–362.
6. Gillet, V.; Willett, P. In Combinatorial Library Design and Evaluation; Ghose, A.; Viswanadhan, V., Eds.; CRC Press: 2001 pp. 379–398.
7. Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors, 1st ed.; Wiley-VCH: Weinheim, Germany, 2000.
8. Meyer, C. Matrix Analysis and Applied Linear Algebra; Society for Industrial and Applied Mathematics, Philadelphia, USA, 2001.
9. Topliss, J.; Edwards, R. J Med Chem 1979, 22, 1238.
10. Mager, P. Med Res Rev 1982, 2, 93.
11. Hubálek, Z. Biol Rev Cambridge Philos Soc 1982, 57, 669.
12. Flower, D. J Chem Inform Comput Sci 1998, 38, 379.
13. Willett, P. Drug Discov Today 2006, 11, 1046.
14. Willett, P.; Barnard, J.; Downs, G. J Chem Inform Comput Sci 1998, 38, 983.
15. François, D. High-Dimensional Data Analysis: Optimal Metrics and Feature Selection. PhD Thesis, Université Catholique de Louvain, Louvain-la-Neuve, Belgium, 2007.
16. François, D.; Wertz, V.; Verleysen, M. IEEE Trans Knowl Data Eng 2007, 19, 873.
17. Fechner, U.; Schneider, G. ChemBioChem 2004, 5, 538.
18. Schneider, P.; Schneider, G. QSAR Comb Sci 2003, 22, 713.
19. Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Angew Chem Int Ed 1999, 38, 2894.
20. Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. J Comput Aided Mol Des 2003, 17, 687.
21. Scott, D.; Thompson, J. In Computer Science and Statistics: Proceedings of the 15th Symposium on the Interface (Interface 1983); Gentle, J., Ed.; North-Holland Publishing Company: Houston, TX, March 17–18, pp. 173–179.
22. Weisgerber, D. J Am Soc Inform Sci 1997, 48, 349.
23. Xue, L.; Godden, J.; Bajorath, J. J Chem Inform Comput Sci 1999, 39, 881.
24. Xue, L.; Godden, J.; Bajorath, J. J Chem Inform Comput Sci 2000, 40, 1227.
25. Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. J Mol Struct 2000, 503, 17.
26. Viswanadhan, V.; Ghose, A.; Revankar, G.; Robins, R. J Chem Inform Comput Sci 1989, 29, 163.
27. Guyon, I.; Elisseeff, A. J Mach Learn Res 2003, 3, 1157.
28. Fodor, I. A survey of dimension reduction techniques. Tech. Rep. UCRL-ID-148494, Lawrence Livermore National Laboratory, Livermore, California, USA, 2002.
29. Hamming, R. Coding and Information Theory; Prentice-Hall, 1980.
30. Abramowitz, M.; Stegun, I. Handbook of Mathematical Functions; Dover, New York, USA, 1972.
31. Beyer, K.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. In Proceedings of the 7th International Conference on Database Theory (ICDT 1999), Jerusalem, Israel, January 10–12, Lecture Notes in Computer Science, New York, USA, vol. 1540, pp. 217–235.
32. Hinneburg, A.; Aggarwal, C.; Keim, D. In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000); Abbadi, A. E.; Brodie, M.; Chakravarthy, S.; Dayal, U.; Kamel, N.; Schlageter, G.; Whang, K.-Y., Eds.; Morgan Kaufmann: Cairo, Egypt; September 10–14, pp. 506–515.
33. Aggarwal, C.; Hinneburg, A.; Keim, D. In Proceedings of the 8th International Conference on Database Theory (ICDT 2001); den Bussche, J. V.; Vianu, V., Eds.; Springer: London, United Kingdom, January 4–6, Lecture Notes in Computer Science, Vol. 1973. pp. 420–434.
34. Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. Numerical Recipes in C., 3rd ed.; Cambridge University Press, New York, USA, 2007.
35. Schneider, G.; Schneider, P.; Renner, S. QSAR Comb Sci 2006, 25, 1162.
36. Xing, E.; Ng, A.; Jordan, M.; Russell, S. In Advances in Neural Information Processing Systems 15 (NIPS 2002); Becker, S.; Thrun, S.; Obermayer, K., Eds.; MIT Press: Cambridge, MA; December 10–12, pp. 505–512.
37. Weinberger, K.; Saul, L. In Proceedings of the 25th International Conference on Machine Learning (ICML 2008); McCallum, A.; Roweis, S., Eds.; Omnipress: Helsinki, Finland; July 5–9, pp. 1160–1167.
38. Weinberger, K.; Blitzer, J.; Saul, L. In Advances in Neural Information Processing Systems 18 (NIPS 2005); Weiss, Y.; Schölkopf, B.; Platt, J., Eds.; MIT Press: Cambridge MA; December 5–8, pp. 1473–1480.
39. Tversky, A. Psychol Rev 1977, 84, 327.