Chapter 2

Molecular Similarity Measures

Gerald M. Maggiora and Veerabahu Shanmugasundaram

Abstract

Molecular similarity is a pervasive concept in chemistry. It is essential to many aspects of chemical reasoning and analysis and is perhaps the fundamental assumption underlying medicinal chemistry. Dissimilarity, the complement of similarity, also plays a major role in a growing number of applications of molecular diversity in combinatorial chemistry, high-throughput screening, and related fields. How molecular information is represented, called the representation problem, is important to the type of molecular similarity analysis (MSA) that can be carried out in any given situation. In this work, four types of mathematical structure are used to represent molecular information: sets, graphs, vectors, and functions. Molecular similarity is a pairwise relationship that induces structure into sets of molecules, giving rise to the concept of chemical space. Although all three concepts – molecular similarity, molecular representation, and chemical space – are treated in this chapter, the emphasis is on molecular similarity measures. Similarity measures, also called similarity coefficients or indices, are functions that map pairs of compatible molecular representations that are of the same mathematical form into real numbers usually, but not always, lying on the unit interval. This chapter presents a somewhat pedagogical discussion of many types of molecular similarity measures, their strengths and limitations, and their relationship to one another. An expanded account of the material on chemical spaces presented in the first edition of this book is also provided. It includes a discussion of the topography of activity landscapes and the role that activity cliffs in these landscapes play in structure–activity studies.

Key words: Molecular similarity, Molecular similarity analyses, Dissimilarity, Activity landscapes

1. Introduction

Similarity is a fundamental concept that has been used since before the time of Aristotle. Even in the sciences, it has been used for more than two centuries [1]. Similarity is subjective and relies upon comparative judgments – there is no absolute standard of similarity, rather "like beauty, it is in the eye of the beholder." Because of this subjectivity, it is difficult to develop methods for unambiguously computing the similarities of large sets of

Jürgen Bajorath (ed.), *Chemoinformatics and Computational Chemical Biology*, Methods in Molecular Biology, vol. 672,
DOI 10.1007/978-1-60761-839-3_2,

molecules [2]. Moreover, there is no absolute standard to compare to so that assessing the validity of any similarity-based method remains subjective; basically, one must rely upon the judgment of experienced scientists. Nevertheless, numerous approaches have been developed over the years to address this difficult but important problem [3–5].

The notion of similarity is fundamental to many aspects of chemical reasoning and analysis; indeed, it is perhaps the fundamental assumption underlying medicinal chemistry, and falls under the general rubric of *molecular similarity analysis* (MSA). Determining the similarity of one "molecular object" to another is basically an exercise in pattern matching – generally called the *matching problem*. The outcome of the exercise is a value, the *similarity measure* that characterizes the degree of matching, association, proximity, resemblance, alignment, or similarity of pairs of molecules as manifested by their "molecular patterns," which are made up of sets of features. The terminology "proximity" is sometimes used in a more general sense to refer to the similarity, dissimilarity, or distance between pairs of molecules. Similarity is generally considered to be a symmetric property, that is "A" is as similar to "B" as "B" is to "A," and most studies are based upon this property. Tversky [6], however, has argued persuasively that certain similarity comparisons are inherently asymmetric. Although his work was directed towards psychology it nonetheless has applicability in studies of molecular similarity. An example will be presented that illustrates the nature of asymmetric similarity and how it can be used to augment the usefulness of the usual symmetric version of similarity (*Cf.* the discussion of Chen and Brown [7]). Willett, Barnard, and Downs [8] presented a comprehensive overview of many of the similarity measures in use today. Their review included a table that summarized the form of the various measures with respect to the type of representation used and should be consulted for further details. Bender and Glen [9] have provided a more recent review of molecular similarity.

Choosing an appropriate feature set and an associated mathematical structure (e.g. set, vector, function, or graph) for handling them is called the *representation problem* and underlies all aspects of MSA. Because similarity is subjective, choosing a feature set depends upon the background of the scientist doing the choosing and to some extent on the problem being addressed. For example, a synthetic organic chemist may focus on the nature of a molecular scaffold and its substituent groups while a physical chemist may be more interested in 3-D shape and electrostatic properties.

Closely allied with the notion of molecular similarity is that of a *chemical space*. Chemical spaces provide a means for conceptualizing and visualizing the molecular similarities of large sets of molecules. A chemical space consists of a set of molecules and a set of associated relations (e.g. similarities, dissimilarities, distances,

etc.) among the molecules that give the space a "structure" [10]. In most chemical spaces, which are coordinate-based, molecules are generally depicted as points. This, however, need not always be the case – sometimes only similarities or "distances" among molecules in the population are known. Nevertheless, this type of pairwise information can be used to construct appropriate coordinate systems using methods such as multi-dimensional scaling (MDS) [11], principal-component analysis (PCA) [12], or nonlinear mapping (NLM) [13] that optimally preserve the information. Coordinate-based chemical spaces can also be partitioned into cells and are usually referred to as cell-based chemical spaces [14]. Each particular type of representation of chemical space has its strengths and weaknesses so that it may be necessary to use multiple types of representations to satisfactorily treat specific problems.

Identifying the appropriate molecular features is crucial in MSA, since the number of potential features is quite large and many contain redundant information. Typical types of molecular features include molecular size, shape, charge distribution, conformation states, and conformational flexibility. In general, only those features deemed relevant or necessary to the matching task at hand are considered. Features are mimicked by any number of descriptors that, ideally, capture the essential characteristics of the features. For example, numerous descriptors of molecular shape exist such as the Jurs shape indices [15] or the Sterimol parameters [16] as well as descriptors of charge distributions such as the venerable Mulliken population analysis [17] or charged partial surface areas, which conveniently incorporate both charge and shape information [18] and descriptors of conformational flexibility such as the Kier molecular flexibility index Φ [19]. Sometimes the term "feature" is used interchangeably with "descriptor." As is seen in the above discussion, features are more general than descriptors, but this distinction is generally not strictly adhered to in most research papers including this one. Other chapters in this work should be consulted for detailed discussion of the many types and flavors of descriptors in use in cheminformatics and chemometrics today.

Similarity measures for assessing the degree of matching between two molecules given a particular representation constitute the main subject matter of this chapter. These measures are functions that map pairs of *compatible* molecular representations (i.e. representations of the same mathematical form) into real numbers usually, but not always, lying on the unit interval. Set, graph, vector, and function-based representations employ a variety of distance and "overlap" measures. Graph-based representations use chemical distance or related graph metrics [20, 21], although numerous graph invariants have been employed as descriptors in vector-based representations [22–24]. All of the

similarity and related measures have at least some idiosyncratic behavior, which can give rise to misleading assessments of similarity or dissimilarity [2]. Similarity measures are sometimes referred to as similarity coefficients or similarity indices and these terminologies will be used somewhat interchangeably in this work.

From the above discussion it is clear that similarity measures provide assessments that are inherently subjective in nature. Thus, the inconsistencies of various measures are not entirely surprising and sometimes can be quite daunting. An interesting approach was developed by Willett's group using a technique called "data fusion" [25]. They showed that values obtained from multiple similarity methods combined using data fusion led to an improvement over similarity-based compound searching using a single similarity method. Subsequent work by Willett's group extended the methodology [26, 27] and provided a detailed account of its conceptual framework [28]. Alternatively, less sophisticated, approaches such as taking the mean of multiple similarity values can also be used.

A brief introduction to the types of molecular representations typically encountered in MSA is presented at the beginning of Subsection 2 followed in Subsection 2.1 by a discussion of similarity measures based upon chemical-graph representations. While graph-based representations are the most familiar to chemists, their use has been somewhat limited in similarity studies due to the difficulty of evaluating the appropriate similarity measures. This section is followed by a discussion of similarity measures based upon finite vector representations, the most ubiquitous types of representations. In these cases, the vector components can be of four types

$$\begin{aligned} &\mathbb{B} \quad \text{Boolean Variables } \{0,1\} \\ &\mathbb{K} \quad \text{Categorical Variables } \{\text{finite, ordered set}\} \\ &\mathbb{N}_o \quad \text{Non - Negative Integer Variables } \{0,1,2,3,\ldots\} \\ &\mathbb{R} \quad \text{Real Variables } \{\text{uncountably infinite set}\} \end{aligned} \tag{1}$$

the first of which called "binary vectors," "bit vectors," or "molecular fingerprints" is by far the most prevalent in applications and is discussed in detail in Subsection 2.2.1. Although the terminology "vector" is used, these objects mathematically are classical sets. Thus, the associated similarity measures are set-based rather than vector-based measures. In addition to the more traditional symmetric similarity measures, a discussion of *asymmetric* similarity measures associated with binary vectors is presented in Subsection 2.2.2.

Vectors whose components are based upon categorical or integer variables are described in Subsection 2.2.3. As was the case for binary vectors, these vectors are also classical sets, and as was the case in the previous subsection, the associated similarity

measures are set-based rather than vector-based. Here, it will also be seen that the form of the set measures are, in some cases, modified from those associated with traditional classical sets.

Subsection 2.3 describes the last class of finite feature vectors, namely those with continuous-valued components, where the components (i.e. features) are usually obtained from computed or experimentally measured properties. An often-overlooked aspect of continuous feature vectors is the inherent non-orthogonality of the basis of the "feature space." The consequences of this are discussed in Subsection 2.3.2. Similarity measures derived from continuous vectors are generally related to Euclidean distances or to cosine or correlation coefficients, all of which are vector-based measures, and are discussed in Subsection 2.3.3.

Essentially, none of the previously discussed approaches deals with the three-dimensionality of molecules. This is dealt with in Subsection 2.4, which describes the application of field-based functions to 3-D molecular similarity. The fields referred to here are related to the steric, electrostatic, and lipophilic properties of molecules and are represented by functions (i.e. "infinite-dimensional vectors"), which are usually taken to be linear combinations of atomic-centered Gaussians. Similarity measures totally analogous to those defined for finite-dimensional, continuous-valued feature vectors (*see* Subsection 2.3.3) also apply here and are treated in Subsection 2.4.2. An often unappreciated issue in 3-D molecular similarity studies is that of *consistent multi-molecule 3-D alignments*, which are discussed in Subsection 2.4.3. Consider the alignment of molecules A and B and that of molecules B and C. Superimposing the aligned pairs using molecule B as a reference induces an alignment of molecules A and C. Now align molecules A and C independently. A consistent multi-molecule alignment is one in which both the induced and independent alignments are essentially the same. As was discussed by Mestres et al. [29], this approach is helpful in identifying "experimentally correct" alignments for a set of reverse transcriptase inhibitors even though the proper conformer of one of the molecules was not in its computed lowest-energy conformation. The role of conformational flexibility in 3-D MSA is discussed in general terms in Subsection 2.4.4. Two general approaches to this problem are described here. One involves the identification of a set of conformational prototypes and the other involves the simultaneous maximization of the similarity measure and minimization of the conformational energy of the molecules being aligned. The former approach is more computationally demanding because it involves $M \times N$ pairwise comparisons, where M and N are the respective numbers of prototype conformations for each pair of molecules. Given that multiple conformations may be important in MSA, how does one determine a similarity value that accounts for multiple conformational states? The discussion in Subsection 2.4.5 suggests an

approach that employs a weighting function based on Boltzmann-like probabilities for each of the conformational states.

Subsection 2.5 provides a brief discussion of molecular dissimilarity, a subject of importance when considering a variety of topics from selecting diverse subsets from a compound collection to the design of diverse combinatorial compound libraries.

The emerging role of *chemical space* in cheminformatics is treated in Subsection 3. It includes a discussion of the dimension of chemical spaces in Subsection 3.1 and a description in Subsection 3.2 of the methods for constructing coordinate-based and coordinate-free chemical spaces, how they can be transformed into one another, and how the usually high-dimension of typical chemical spaces can be reduced in order to facilitate visualization and analysis. The closely related subject of *activity cliffs* and the topography of activity landscapes are discussed in Subsection 3.3. How the information contained in activity landscapes, which are inherently of high-dimension, can be portrayed in lower dimensions is discussed. Emphasis is placed on the use of structure–activity similarity (SAS) maps, although several other recent approaches to this problem are described. An information-theoretic analysis of SAS maps is also presented. Subsection 3 ends with a somewhat detailed description of a general similarity-based approach for representing chemical spaces (*see* Subsection 3.4). The method explicitly accounts for the inherent non-orthogonality of vector representations of chemical space. Unlike some of the vector-like methods described earlier, this method employs "molecular vectors" that actually live in a linear vector space.

The present work is not intended as a comprehensive review of the similarity literature. Rather, it is intended to provide an integrated and somewhat pedagogical discussion of many of the simple, complex, and confounding issues confronting scientists using the concept of molecular similarity in their work.

2. Molecular Representations and Their Similarity Measures

How the structural information in molecules is represented is crucial to the types of "chemical questions" that can be asked and answered. This is certainly true in MSA where different representations and their corresponding similarity measures can lead to dramatically different results [2]. Four types of mathematical objects are typically used to represent molecules – sets, graphs, vectors, and functions. Sets are the most general objects and basically underlie the other three and are useful in their own right as will be seen below. Because of their importance a brief introduction to sets, employing a more powerful but less familiar

notation than that typically used, is provided in the Appendix (*see* Subsection 5).

Typically, chemists represent molecules as "chemical graphs" [30], which are closely related to the types of graphs dealt with by mathematicians in the field of graph theory [31]. Most chemical graphs describe the nature of the atoms and how they are bonded. Thus, chemical graphs are sometimes said to provide a 2-D representation of molecules. They do not typically contain information on the essential 3-D features of molecules, although chemical graphs have been defined that do capture some of this information [32]. Three-dimensional structures are also used extensively, especially now that numerous computer programs have been developed for their computation and display.

While chemical graphs provide a powerful and intuitive metaphor for understanding many aspects of chemistry, they nevertheless have their limitations especially when dealing with questions of interest in chemometrics and cheminformatics. In these fields, molecular information is typically represented by *feature vectors*, where each component corresponds to a "local" or "global" feature or property of a molecule usually represented by one of a number of possible descriptors associated with the chosen feature. Local features include molecular fragments ("substructures"), potential pharmacophores [33], various topological indices [34], and partial atomic charges, to name a few. Global features include properties such as molecular weight, logP, polar surface area, various BCUTs [35], and volume. It is well to point out here that use of the term "vector" is not strictly correct. For example, in Subsection 2.2 on Discrete-Valued Feature Vectors the "bit vectors" used to depict molecular fingerprints are actually classical sets or multisets that do not strictly behave according to the rules of linear vector spaces [36]. For example, bit vectors $\mathbf{v}_A$ and $\mathbf{v}_B$ do not satisfy the additive ("+") and scalar multiplicative (" $\cdot$ ") properties associated with vectors residing in linear vector spaces, i.e. $\mathbf{v}_C \neq a \cdot \mathbf{v}_A + b \cdot \mathbf{v}_B$, where a and b are any real scalars. As discussed in the sequel, some classes of continuous-valued vectors (*see* Subsection 2.3) such as BCUTS are also not vectors in the strict mathematical sense since they do not strictly obey the additive property of vectors (e.g. the sum of two BCUTS may lie outside of the BCUT chemical space).

More recently, with the significant increases in computer power even on desktop PCs, methods for *directly matching* 3-D features of molecules have become more prevalent. Features here generally refer to various types of molecular fields, some such as electron density ("steric") and electrostatic-potential fields are derived from fundamental physics [37, 38] while others such as lipophilic potential fields [39] are constructed in an ad hoc manner. Molecular fields are typically represented as continuous functions. As is the case for discrete and continuous vectors noted in

the previous paragraph, such functions also may not strictly satisfy the axioms of linear function spaces [40]. However, this does not preclude their usefulness in determining 3-D molecular similarities. Discrete fields have also been used [41], albeit somewhat less frequently except in the case of the many CoMFA-based studies [42].

In cheminformatics, similarities typically are taken to be real numbers that lie on the unit interval [0,1]. However, since similarity is an inherently "fuzzy" concept, it may be appropriate to take a fuzzier view. Bandemere and Näther [43] have written extensively on an approach that treats similarity as a fuzzy number [44]. Fuzzy numbers can be conceptualized as Gaussian functions or, more commonly as "teepee-like" functions, with maximum value unity, the "width" of the function being associated with the "degree of fuzziness" of the numbers. While this is a conceptually powerful approach, it suffers many of the problems associated with fuzzy numbers. For example, it two fuzzy numbers overlap significantly determining how much larger one fuzzy number is to the other can become difficult [44]. Thus, for example, comparing how similar two molecules are to a given molecule can become a significant problem. Nevertheless, investigating the realm of fuzzy similarities may prove to be a suitable approach to similarity, but further work needs to be done before any reasonable conclusion can be obtained as to its usefulness in cheminformatics.

2.1. Chemical Graphs

Chemical graphs are ubiquitous in chemistry. A chemical graph, G_k, can be defined as an ordered triple of sets

$$G_k = (V_k,\ E_k,\ L_k), \tag{2}$$

where V_k is a set of n vertices ("atoms") and $V_k(x_i)$ is an *indicator* or *characteristic function* with values $V_k(x_i) = \{0, 1\}$ that designates the respective absence or presence of a given vertex

$$V_k = \{V_k(x_1), V_k(x_2), ..., V_k(x_n)\}. \tag{3}$$

Alternatively, V_k can be given, in more familiar notation, by

$$V_k = \{v_{k,k_1}, v_{k,k_2}, ..., v_{k,k_\ell}\} \tag{4}$$

where only the ℓ vertices for which $V(x_i) = 1$ are explicitly designated (*See* Appendix Subsection 5 for notational details). The edge set, E_k, can be written in analogous fashion,

$$E_k = \{e_{k,k_1}, e_{k,k_2}, ..., e_{k,k_m}\}, \tag{5}$$

where the m elements of the set are the edges ("bonds") between vertices ("atoms"), and each edge is associated with an unordered pair of vertices, $e_{k,k_i} = \{v_{k,k_p}, v_{k,k_q}\}$. The label set, L_k, is a set of r symbols

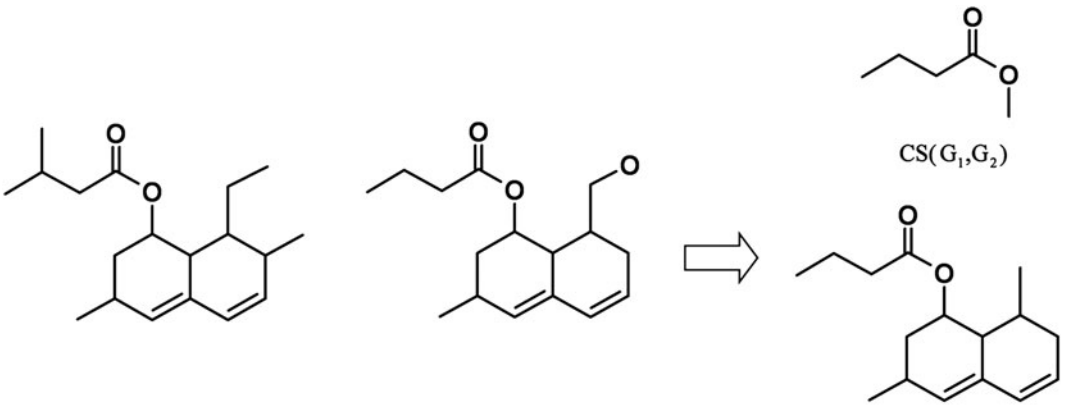
$$L_k = \{\ell_{k,k_1}, \ell_{k,k_2}, ..., \ell_{k,k_r}\} \tag{6}$$

that label each vertex ("atom") and/or edge ("bond"). Typical atom labels include hydrogen ("H"), carbon ("C"), nitrogen ("N"), and oxygen ("O"); typical bond labels include single ("s"), double ("d"), triple ("t"), and aromatic ("ar"), but other possibilities exist. Whatever symbol set is chosen will depend to some degree on the nature of the problem being addressed. In most cheminformatics applications, *hydrogen suppressed* chemical graphs are used, which are obtained by deleting all of the hydrogen atoms. Fig. 1 depicts an example of two hydrogen-suppressed chemical graphs, G_1 and G_2, which are clearly related to a chemist's 2-D representation of a molecule. Chemical graphs of 3-D molecular structures are described by Raymond and Willett [32], but their use has been much more limited.

The notion of a subgraph is also important. If G'_k is a subgraph of G_k, written $G'_k \subseteq G_k$, then

$$G'_k \subseteq G_k \;\Rightarrow\; V'_k \subseteq V_k \text{ and } E'_k \subseteq E_k, \tag{7}$$

that is the vertex and edge sets V'_k and E'_k associated with the subgraph, G'_k, are subsets of the corresponding vertex and edge sets V_k and E_k of the graph, G_k. Many operations defined on sets



$$S_{Tan}(G_i,G_j) = \frac{|G_i \cap G_j|_E}{|G_i \cup G_j|_E} = \frac{|MCS(G_i,G_j)|_E}{|G_i|_E + |G_j|_E - |MCS(G_i,G_j)|_E} = \frac{19}{22+20-19} = 0.83$$

$$d(G_i,G_j) = |G_i|_E + |G_j|_E - 2|MCS(G_i,G_j)|_E = 22+20-2(19) = 4$$

Fig. 1. An example of two hydrogen-suppressed graphs G1, G2 and a common substructure CS(G1,G2) and the maximum common substructure MCS(G1,G2) are shown above. The Tanimoto similarity index and the distance between the two chemical graphs are computed below.

can also be defined on graphs. One such operation is the norm or cardinality of a graph,

$$|G_k| = |V_k| + |E_k| \tag{8}$$

which is a measure of the "size" of the graph. Another measure is the *edge norm* that is given by

$$|G_k|_{\mathrm{E}} = |E_k|, \tag{9}$$

where the subscript E explicitly denotes that the cardinality refers only to the edges ("bonds") of the graph. For the two chemical graphs depicted in Fig. 1, $|G_1|_{\mathrm{E}} = 22$ and $|G_2|_{\mathrm{E}} = 20$. Note that only the number of bonds and not their multiplicities (e.g. single, double, etc.) are considered here. However, many other possibilities exist, and their use will depend on the problem being addressed [20].

A key concept in the assessment of molecular similarity based upon chemical graphs is that of a *maximum common substructure*, $\mathrm{MCS}(G_i, G_j)$, of two chemical graphs, which derives from the concept of maximum common subgraph employed in mathematical graph theory. There are several possible forms of MCS [21, 32]. Here, we will focus on what is usually called the maximum common edge substructure, which is closest to what chemists perceive as "chemically meaningful" substructures [45], but we will retain the simpler and more common nomenclature MCS. A common (edge) substructure (CS) of two chemical graphs is given by

$$\mathrm{CS}(G_i, G_j)_{k,\ell} = E_i^k \cap E_j^\ell = E_i^k = E_j^\ell, \tag{10}$$

where E_i^k and E_j^ℓ are subsets of their respective edge sets, $E_i^k \subseteq E_i$ and $E_j^\ell \subseteq E_j$, and are equivalent. Thus, the intersection (or union) of these two equivalent subsets is equal to the sets themselves. As there are numerous such common substructures, $\mathrm{CS}(G_i, G_j)_{k,\ell}$, $k, \ell = 1, 2, 3, \ldots\ldots$, determining the MCS between two chemical graphs is equivalent to determining the edge intersection-set of maximum cardinality, that is

$$\begin{aligned} \mathrm{MCS}(G_i, G_j) &= \mathrm{CS}(G_i, G_j)_{p,q} \text{ such that } \left|\mathrm{CS}(G_i, G_j)_{p,q}\right|_{\mathrm{E}} \\ &= \max_{k,\ell} \left|\mathrm{CS}(G_i, G_j)_{k,\ell}\right|_{\mathrm{E}} \end{aligned} \tag{11}$$

Thus,

$$G_i \cap G_j \equiv \mathrm{MCS}(G_i, G_j), \tag{12}$$

that is the MCS is equivalent to "graph intersection," which is equivalent to the maximum number of edges in common between the two molecules. Note that multiple solutions may exist and that some of the solutions could involve disconnected graphs.

However, to obtain "chemically meaningful" results only *connected* MCS's are usually considered.

The edge cardinality of the intersection and union of two chemical graphs is given, respectively, by

$$\left|G_i \cap G_j\right|_{\mathrm{E}} = \left|\mathrm{MCS}(G_i, G_j)\right| \tag{13}$$

and

$$\left|G_i \cup G_j\right|_{\mathrm{E}} = \left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - \left|\mathrm{MCS}(G_i, G_j)\right|. \tag{14}$$

These two expressions form the basis for several measures such as Tanimoto similarity (*see* Subsection 2.2 for an extensive discussion)

$$S_{\mathrm{Tan}}(G_i,G_j) = \frac{\left|G_i \cap G_j\right|_{\mathrm{E}}}{\left|G_i \cup G_j\right|_{\mathrm{E}}} = \frac{\left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}}{\left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - \left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}} \tag{15}$$

and the distance between two chemical graphs

$$d(G_i,G_j) = \left|G_i\right|_{\mathrm{E}} + \left|G_j\right|_{\mathrm{E}} - 2\left|\mathrm{MCS}(G_i,G_j)\right|_{\mathrm{E}}. \tag{16}$$

The edge cardinality is explicitly designated in Eqs. (11) and (13)–(16) in order to emphasize that a particular norm has been chosen. Equation (15) is the graph-theoretical analog of the well-known Tanimoto similarity index (*see* Eq. (21)), which is symmetric and bounded by zero and unity. Equation (16) corresponds to the distance between two graphs [46], which is the number of bonds that are not in common in the two molecules depicted by G_i and G_j. Another distance measure called "chemical distance" is similar to that given in Eq. (16) except that lone-pair electrons are explicitly accounted for [21]. The Tanimoto similarity index of the two chemical graphs in Fig. 1 and the distance between them are given by $S_{\mathrm{Tan}}(G_i,G_j) = 0.83$ and $d(G_i,G_j) = 4$, respectively.

A similarity index called "subsimilarity," which is short for substructure similarity, has been developed by developed by Hagadone [47]. In form it is identical to one of the family of asymmetric similarity indices developed by Tversky [6] that is discussed in Subsection 2.2.2,

$$S_{\mathrm{Tev}}(G_{\mathrm{Q}},G_{\mathrm{T}}) = \frac{\left|G_{\mathrm{Q}} \cap G_{\mathrm{T}}\right|_{\mathrm{E}}}{\left|G_{\mathrm{Q}}\right|_{\mathrm{E}}} = \frac{\left|\mathrm{MCS}(G_{\mathrm{Q}},G_{\mathrm{T}})\right|_{\mathrm{E}}}{\left|G_{\mathrm{Q}}\right|_{\mathrm{E}}}, \tag{17}$$

where $G_{\mathrm{Q}}$is the substructure query and G_{T} is a target molecule. In contrast to $S_{\mathrm{Tan}}(G_i,G_j)$, $S_{\mathrm{Tev}}(G_i,G_j)$ is not symmetric, although zero and unity also bound it.

While chemical graphs are intuitive to those trained in the chemical sciences, they have not been widely used in MSA primarily because of the computational demands brought on by the need to compute $\mathrm{MCS}(G_i,G_j)$, which for large complex systems can be quite daunting. Approximate algorithms do exist, however

[32, 47] and with the ever-increasing power of computers the use of graph-based similarity may become more prevalent in the future. Interestingly, there is a close analogy between determination of the MCS and alignment of the 3-D molecular fields of molecules (*see* Subsection 2.4) except that in the former the optimization is discrete while in the latter it is continuous.

2.2. Discrete-Valued Feature Vectors

The components of discrete feature vectors may indicate the presence or absence of a feature, the number of occurrences of a feature, or a finite set of binned values such as would be found in an ordered, categorical variable.

2.2.1. Binary-Valued Feature Vectors

Each component of an n-component binary feature vector, also called *bit vectors* or *molecular fingerprints*,

$$\mathbf{v}_A = (v_A(x_1), v_A(x_2), ..., v_A(x_k), ..., v_A(x_n)) \tag{18}$$

indicates the presence or absence of a given feature, x_k, that is

$$v_A(x_k) = \begin{cases} 1 & \text{Feature present} \\ 0 & \text{Feature absent} \end{cases}. \tag{19}$$

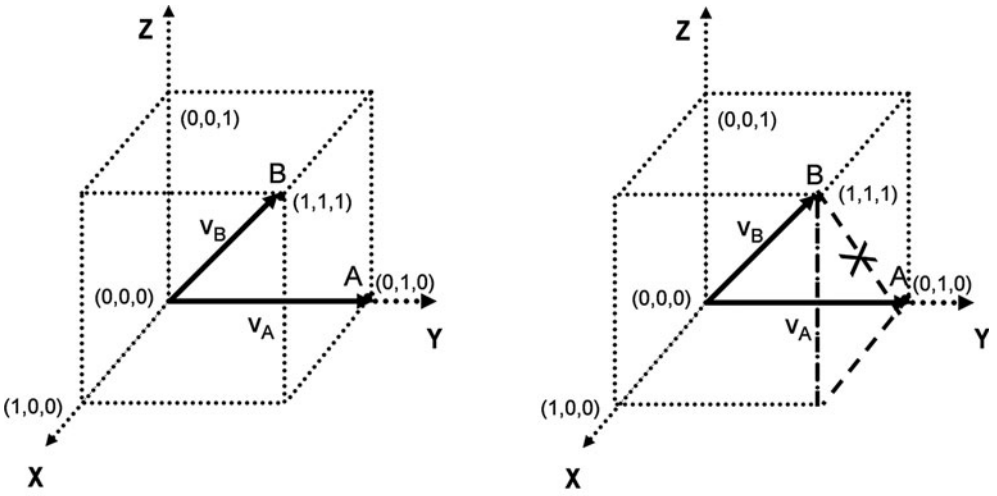
A wide variety of features have been used in bit vectors. These include molecular fragments, 3-D "potential pharmacophores," atom pairs, 2-D pharmacophores, topological torsions, and variety of topological indices to name a few.

Binary feature vectors are completely equivalent to sets (*see* the Appendix in Subsection 5 for further discussion). Care must be exercised when using them to ensure that appropriate mathematical operations are carried out. The number of components in a bit vector is usually quite large, normally $n >> 100$. In some cases n can be orders of magnitude larger, sometimes exceeding a million components [33, 48]. Bit vectors of this size are not handled directly since many of the components are zero, and methods such as hashing [49] are used to reduce the size of the stored information.

Bit vectors live in an n-dimensional, discrete hypercubic space, where each vertex of the hypercube corresponds to a set. Figure 2 provides an example of sets with three elements. Distances between two bit vectors, $\mathbf{v}_A$ and $\mathbf{v}_B$, measured in this space correspond to Hamming distances, which are based upon the city-block ℓ_1 metric

$$d_{Ham}(\mathbf{v}_A, \mathbf{v}_B) = |\mathbf{v}_A - \mathbf{v}_B| = \sum_{k=1}^{n} |v_A(x_k) - v_B(x_k)|. \tag{20}$$

Since these vectors live in an n-dimensional hypercubic space, the use of non-integer distance measures is inappropriate, although in this special case the square of the Euclidean distance is equal to the Hamming distance.



$$d_{\text{Ham}}(\mathbf{v}_A, \mathbf{v}_B) = |\mathbf{v}_A - \mathbf{v}_B| = \sum_{k=1}^{n} |v_A(x_k) - v_B(x_k)| = [1-0] + [1-1] + [1-0] = 2$$

Fig. 2. Distance between two binary-valued feature vectors $\mathbf{v}_A$ and $\mathbf{v}_B$ is not given by the Euclidean distance but the Hamming distance between the two.

The most widely used similarity measure by far is the Tanimoto similarity coefficient S_{Tan}, which is given in set-theoretic language as (*Cf.* Eq. (15) for the graph-theoretical case)

$$S_{\text{Tan}}(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{21}$$

Using the explicit expressions for set cardinality, intersection and union given by Eqs. (116, 111 and 112), respectively, in the Appendix, Eq. (21) becomes

$$S_{\text{Tan}}(A, B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k \max[A(x_k), B(x_k)]}. \tag{22}$$

By changing the form of the denominator (*see* Eqs. (120) and (121)), S_{Tan} is also given by

$$\begin{aligned} S_{\text{Tan}}(A, B) &= \frac{|A \cap B|}{|A - B| + |B - A| + |A \cap B|}, \\ &= \frac{a}{a + b + c} \end{aligned} \tag{23}$$

where

$$\begin{aligned} a &= |A \cap B| \quad \text{Number of features common to } A \text{ and } B \\ b &= |A - B| \quad \text{Number of features common to } A \text{ but not to } B. \\ c &= |B - A| \quad \text{Number of features common to } B \text{ but not to } A \end{aligned} \tag{24}$$

The Tanimoto similarity coefficient is symmetric,

$$S_{\mathrm{Tan}}(A,\ B) = S_{\mathrm{Tan}}(B,\ A), \tag{25}$$

as are most of the similarity coefficients in use today, and is bounded by zero and unity,

$$0 \leqslant S_{\mathrm{Tan}}(A,\ B) \leqslant 1. \tag{26}$$

From the form of these equations it can be seen that the method is biased when there is a great disparity in the size of the two molecules being compared. Consider, for example, the case when $|Q| << |T|$, where Q is a query molecule and T is a target molecule that could be obtained in a similarity search. If Q is much smaller than T, $|Q \cup T| \approx |T|$, and since $|Q| \leqslant |Q \cap T|$, it follows that $S_{\mathrm{Tan}}(Q, T) \approx |Q|/|T|$. A consequence of this relationship is that in similarity-based searching Q will tend to recover other small molecules, T, since as T gets larger S_{Tan} becomes smaller in value, which works against the selection of larger molecules in the search. This is not generally a problem except in cases where a substructure of a large target molecule is quite similar to the smaller query molecule. If the query were biologically active, the larger target molecule containing a similar substructure to the query, which is bioactive, would be missed. The same holds true for a large molecule query that is it will tend to recover larger molecules. Thus, molecules with a strong *substructural relationship* to the query molecule will likely be missed, but this could be important in drug design as the substructure may contain the key atoms of the pharmacophore. As will be seen in the next section, the use of an asymmetric similarity measure can compensate for this to some degree. The above argument carries through completely to the case of chemical-graph-based similarity indices (*see* Subsection 2.1).

A number of other similarity indices are in use today. The recent work by Willett, Barnard, and Downs [8] should be consulted for examples of many of them including a comprehensive discussion of their properties.

2.2.2. Asymmetric Similarity Indices

Most similarity measures for binary-valued feature vectors in use today are symmetric, Tversky [6], however, has defined an infinite family of *asymmetric* measures

$$S_{\mathrm{Tve}}(A, B) = \frac{|A \cap B|}{\alpha|A - B| + \beta|B - A| + |A \cap B|}, \tag{27}$$

where $\alpha, \beta \geqslant 0$. This generalizes the typical symmetric Tanimoto similarity measure given in Eq. (23), which obtains when $\alpha = \beta = 1$. For all other values of α and β, $S_{\mathrm{Tve}}(A, B)$is asymmetric, that is $S_{\mathrm{Tve}}(A, B) \neq S_{\mathrm{Tve}}(B, A)$. Only the two extreme forms will,

however, be considered here, namely those when $\alpha = 1$ and $\beta = 0$ and $\alpha = 0$ and $\beta = 1$. Their set-theoretic forms are given by

$$S^*_{\text{Tve}}(A, B) = \frac{|A \cap B|}{|A - B| + |A \cap B|} = \frac{|A \cap B|}{|A|} \quad \text{Fraction of } A \text{ similar to } B \tag{28}$$

$$S^*_{\text{Tve}}(B, A) = \frac{|A \cap B|}{|B - A| + |A \cap B|} = \frac{|A \cap B|}{|B|} \quad \text{Fraction of } B \text{ similar } A \tag{29}$$

Using Eqs. (111) and (116) both of the above equations can be written in a form similar to that for S_{Tan} given in Eq. (22). For example, Eq. (28) becomes

$$S^*_{\text{Tve}}(A, B) = \frac{\sum_k \min[A(x_k), B(x_k)]}{\sum_k A(x_k)}. \tag{30}$$

In analogy to Eq. (23), the asymmetric similarity indices are given, respectively, by

$$S^*_{\text{Tve}}(A, B) = \frac{a}{a + b} \text{ and } S^*_{\text{Tve}}(B, A) = \frac{a}{a + c}. \tag{31}$$

As was the case for the symmetric similarity coefficient

$$0 \leqslant S^*_{\text{Tve}}(A, B), S^*_{\text{Tve}}(B, A) \leqslant 1, \tag{32}$$

although $S^*_{\text{Tev}}(A, B) \neq S^*_{\text{Tev}}(B, A)$, in general. If $|A| < |B| \Rightarrow S^*_{\text{Tev}}(A, B) > S^*_{\text{Tev}}(B,A)$.

Asymmetric similarity can provide some benefits in similarity searches not afforded by its symmetric competitors. For example, consider as in Subsection 2.2.1, the query and target molecules, Q and T, respectively, and the asymmetric similarity coefficients given in Eqs. (28) and (29). If Q is relatively "small," (*N.B.* "small" and "large" are used here refer to the size of the set and not to the size of the corresponding molecule) that is if $|Q| << |T|$, then target molecules for which Q is an approximate subset will be selected using Eq. (28), that is

$$S^*_{\text{Tve}}(Q, T) = \frac{|Q \cap T|}{|Q|} \Rightarrow 1 \text{ as } Q \cap T \Rightarrow Q. \tag{33}$$

This result is approximately independent of the size of T given that Q is an approximate subset of T. A comparable selection of molecules would not be obtained using the symmetric similarity coefficient in Eq. (21) or the asymmetric similarity coefficient given by Eq. (29) since as the target molecule increased in size the denominator would reduce the overall similarity values

making selection less likely. If, on the other hand, Q is a relatively "large," that is if $|Q|>>|T|$, then using the lower expression for asymmetric similarity in Eq. (29) will produce similar results

$$S^{*}_{\mathrm{Tve}}(T, Q) = \frac{|Q \cap T|}{|T|} \Rightarrow 1 \text{ as } Q \cap T \Rightarrow T \tag{34}$$

except that the target molecules retrieved will be smaller than Q and will also be approximate subsets of Q. An example of this is shown in Figs. 3 and 4.

The "extreme" forms, but not the intermediate forms, of asymmetric similarity defined by Tversky [6] given in Eqs. (28) and (29) can be transformed into two symmetric measures by taking the maximum and minimum of the set cardinalities in the denominators of the two equations. The forms of these equations are obtained in analogy to those developed by Petke [41] for vectors and field-based functions (*see* Subsections 2.3 and 2.4 for further details):

$$S_{\mathrm{Pet}_{\max}}(A,\ B) = \frac{|A \cap B|}{\max(|A|, |B|)} \tag{35}$$

and

$$S_{\mathrm{Pet}_{\min}}(A,\ B) = \frac{|A \cap B|}{\min(|A|, |B|)}. \tag{36}$$

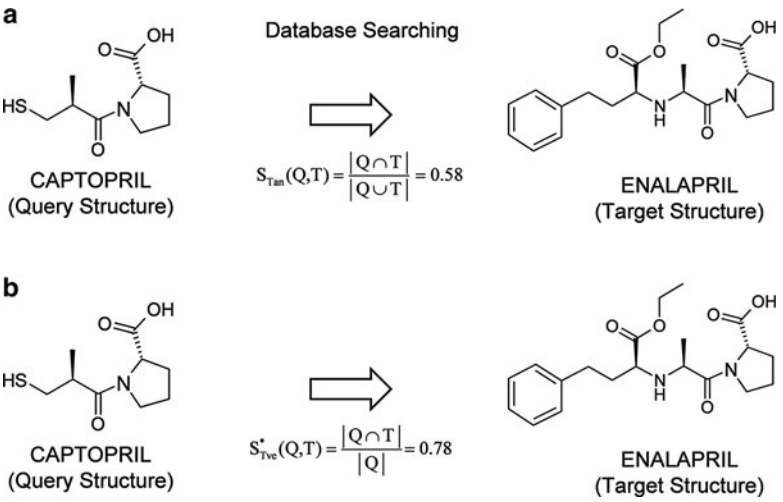


Fig. 3. Asymmetric similarity-based searching might provide some benefits not afforded by symmetric similarity-based searching. (**a**) Database searching using ISIS keys and symmetric (Tanimoto) similarity will not yield enalapril as a "database hit" because the similarity value (0.58) is too low. (**b**) In contrast, database searching using ISIS keys and asymmetric (Tversky) similarity could yield enalapril as a "database hit" because the asymmetric similarity value (0.78) is considerably larger than the corresponding symmetric one (0.58). This illustrates that small query molecules are more likely to retrieve larger target molecules in similarity searches based upon asymmetric rather than symmetric similarity indices.

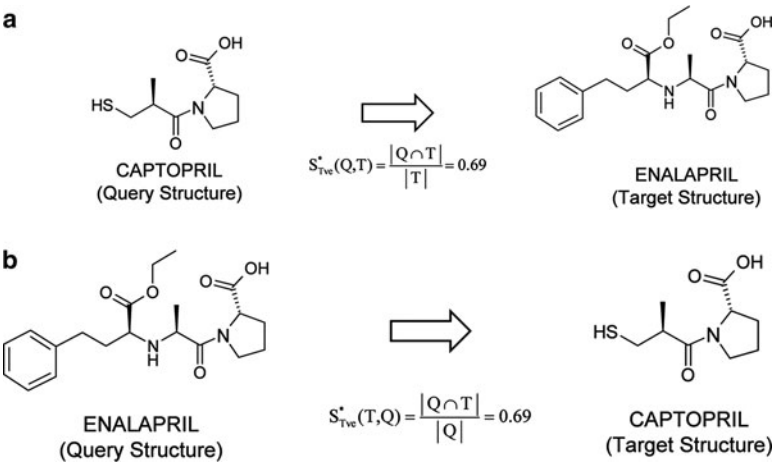


Fig. 4. (**a**) The other asymmetric (Tversky) similarity has a value of 0.69. Exchanging the roles of the query (Q) and target (T) molecules, i.e., Q(captopril) ⇒ T(enalapril) and T(enalapril) ⇒ Q(captopril), gives (**b**), which shows that large query molecules are more likely to retrieve smaller target molecules in similarity searches based upon asymmetric rather than symmetric similarity since the values of the corresponding indices are 0.69 and 0.58, respectively.

As is the case for asymmetric similarity indices, both $S_{\mathrm{Pet}_{\max}}(A,B)$ and $S_{\mathrm{Pet}_{\min}}(A,B)$are bounded by zero and unity, but are ordered with respect to each other and with respect to Tanimoto similarity, that is

$$0 \leqslant S_{\mathrm{Pet}_{\max}}(A,B) \leqslant S_{\mathrm{Tan}}(A,B) \leqslant S_{\mathrm{Pet}_{\min}}(A,B) \leqslant 1. \qquad (37)$$

2.2.3. Integer- and Categorical-Valued Feature Vectors

Feature vectors with integer- or categorical-valued components are identical in form to binary-valued vectors (*see* Eq. (18)). In contrast, however, each component takes on a finite number of values

$$v(x_k) = \begin{cases} \text{Finite, Ordered Set of Non - Negative Integers} \\ \text{Finite, Ordered Set of Values} \end{cases} \qquad (38)$$

In the integer case, these values usually refer to the frequency of occurrence of a given feature such as, for example, a molecular fragment. In the categorical case the values may refer to a binned variable. In both cases, the vectors live in discrete, lattice-like "hyper-rectangular" spaces, which are generalizations of the hypercubic spaces inhabited by bit vectors. Such spaces can also be described by multisets [50], but this formalism will not be used in this work.

Ideally, distances in these spaces should be based upon an $\ell_1$or city-block metric (*see* Eq. (20)) and not the ℓ_2 or Euclidean metric typically used in many applications. The reason for this are the

same as those discussed in Subsection 2.2.1 for binary vectors. Set-based similarity measures can be adapted from those based on bit vectors using a formula borrowed from fuzzy set theory [51, 52]. For example, the Tanimoto similarity coefficient becomes

$$S_{\mathrm{Tan}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) = \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k \max[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]} \tag{39}$$

As noted in Klir and Yuan [51] there are many possible denominators that can be used in place of $|A \cup B|$, each of which gives rise to a different similarity measure.

The asymmetric similarity coefficients become, in an analogous fashion [see Eqs. (30)]

$$\begin{aligned} S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) &= \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k v_{\mathrm{A}}(x_k)} \\ S^{*}_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) &= \frac{\sum_k \min[v_{\mathrm{A}}(x_k),\, v_{\mathrm{B}}(x_k)]}{\sum_k v_{\mathrm{B}}(x_k)} \end{aligned} \tag{40}$$

As was the case in the previous section for bit vectors, it can be shown that the similarity coefficients defined here are also bounded,

$$0 \leqslant S_{\mathrm{Tan}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}) \leqslant S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{A}},\mathbf{v}_{\mathrm{B}}),\ S_{\mathrm{Tve}}(\mathbf{v}_{\mathrm{B}},\mathbf{v}_{\mathrm{A}}) \leqslant 1. \tag{41}$$

in the case of non-negative integer-valued vector components. Other modifications are needed to accommodate non-integer values. Maggiora et al. [53], have discussed this issue, in general, for the case of field-based continuous functions, but their work also applies to "vectors" such as those described here.

In a methodology call holographic QSAR [54], integer-valued vectors are employed to characterize the frequency of occurrence of molecular fragments. The vectors are not, however, used in their "native" form but rather are folded into a smaller vector by hashing. Schneider et al. [55] have also used integer-valued vectors to characterize what they call 2-D pharmacophores.

Integer- and categorical-valued vectors can be converted into equivalent binary vectors by augmenting the components of a typical bit vector as shown in Fig. 5. The process is straightforward for integer-valued variables. Bajorath and co-workers [56] have developed a novel binning approach for variables with continuous values, basically converting them into categorical variables. Once the mapping to the augmented bit vector has been completed all of the usual bit-vector-based similarity measures (*see* Subsection 2.2.2 for further discussion) can be applied.

There are many other expressions for similarity that can be used for integer- and categorical-valued vectors. Again, the

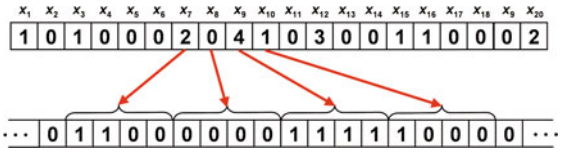


Fig. 5. In the scheme shown above, a 20-bit integer-valued vector (maximum integer value for each bit is 4) is converted into a 80-bit binary vector by converting each integer bit into a binary bit of 4-bit length. {0 = 0000; 1 = 1000; 2 = 1100; 3 = 1110; 4 = 1111}.

comprehensive discussion provided by Willett et al should be consulted for additional details [8]. Many of the features of discrete vector-based representations do not capture all of the relevant 3-D information in any substantive way, although they do capture some 3-D information indirectly, and this is why some feature vector procedures are referred to as "2.5-D" methods.

2.3. Continuous-Valued Feature Vectors

Vectors whose components have continuous values correspond to the more "traditional" types of vectors found in the physical sciences. They are of identical form to the discrete-valued vectors (*see* Eq. (18)) except that the components, $v_A(x_k)$, are continuous valued. In cheminformatics, however, the nature of the components is considerably different from those typically found in physics. For example, physiochemical properties, such as logP, solubility, melting point, molecular volume, Hammett $\sigma\rho$ parameters, surface charge, etc., as well as other descriptors derived explicitly for the purpose, such as BCUTs [35], have been routinely used. The use of continuous-valued vectors is usually confined to relatively low-dimensional chemical spaces, generally of less than ten dimensions (*see* Subsection 3 for further discussion). This is in sharp contrast to those discussed in the previous sections, where the dimensions are generally considerably larger.

Although it is ubiquitous in cheminformatics applications, the term vector should be used with caution as vectors are properly the objects of vector or affine spaces, and hence, must satisfy the axioms of these spaces. For example, vectors in BCUT chemical spaces do not form a vector space since the sum of two BCUT vectors may not lie in the space [35]. However, as long as this rather fine distinction is borne in mind significant problems should not arise, and the term vector, taken in its broadest if not strictest mathematical sense, will be used here. For a more general but brief discussion of vectors see the presentation of Euclidean vectors in Wikipedia [57].

2.3.1. Property-Based Continuous-Valued Feature Vectors

The components of most continuous-valued feature vectors are based on a variety of molecular properties such as solubilities, logPs, melting points, polar surface areas, molecular volumes,

various shape indices, and BCUTs, which are related to the charge, polarizability, and hydrogen bonding properties of molecules. Since these properties have a wide range of values they are typically scaled using the usual "z-transform" $z_i = (x_i - \bar{x})/\sigma_x$ favored by statisticians, where $\bar{x}$ is the average property-value and σ_x^2 is its variance (*N.B.* that this transformation is not strictly appropriate for multi-modal data). Other transforms have also been used; one of the most popular is $x'_i = (x_i - x_{\min})/(x_{\max} - x_{\min})$., where the values of the property, x_i, are mapped into the unit interval [0,1]. Simple scaling can be used to expand or contract the unit interval if desired.

An advantage of the z-transform is that it establishes a well-defined point of reference for the property-based vectors (the mean) as well as scaling the values of all of the variables to unit variance. BCUTs have a more complicated scaling, and the paper by Pearlman and Smith [35] should be consulted for further details. Since distances between vectors are invariant to the origin of the coordinate system, mean centering does not affect the result. However, the transformations used in all of the above procedures involve some form of scaling, and thus distances are not preserved between the original and scaled coordinate systems. Care must be exercised in the case of cosine similarity indices between vectors since they are both origin and scale dependent.

2.3.2. Inherent Non-orthogonality of Descriptor Coordinate Systems

An often-overlooked issue is the *inherent non-orthogonality* of coordinate systems used to portray data points. Almost universally a Euclidean coordinate system is used. This assumes that the original *variables* are orthogonal, that is are uncorrelated, when it is well known that this is generally not the case. Typically, PCA [12] is performed to generate a putative orthogonal coordinate system each of whose axes correspond to directions of maximum variance in the transformed space. This, however, is not quite correct. Since an orthogonal similarity transformation is used to carry out the PCA, and since such transformations rigidly rotate the original coordinate system, the angles among the coordinate vectors are unchanged. By exactly reversing the rigid rotation of the orthogonal principle-component coordinate system ones regenerates the original coordinate system, which is thus seen to be orthogonal. This clearly contradicts the general observation that most variables used in practice tend to be statistically correlated, that is are non-orthogonal. Importantly, even when the variables are properly uncorrelated this does not mean that they are necessarily *statistically independent* [58]. To correctly handle such correlated variables one must first orthogonalize the original variables, and then perform PCA to orient the orthogonal coordinate system along directions of maximum variance of the data points. This is rarely done in current practice, but what are the consequences of not doing this? As is well known from the theory

of tensors [59] both distances and angles between data vectors are affected by the angles between the coordinate axes (*Cf.* the discussion presented in Subsection 3.1.3 and in the paper by Raghavendra and Maggiora [95]). Conclusions drawn using, for example, either cosine similarity indices or distances will be affected *quantitatively* but not *qualitatively*. This is a manifestation of the fact that the topology (i.e. neighborhood relationships) of the space is preserved but its geometry (i.e. distances and angles) is not. The consequences of this are the following. The order of nearest-neighbors from a given reference molecule in a chemical space (*see* Subsection 3 for further details) will remain unchanged but the magnitude of their distances from the reference molecule will change. Thus, if one is only interested in, say, obtaining the 50 most similar molecules to a given reference molecule nothing will change by modifying the angles of the coordinate axes. If, one the other hand, one is interested in finding all molecules with similarities greater than or equal to, say, 0.85 with respect to that reference molecule, the results obtained will change since they depend on the angles of the coordinate vectors.

In many cases, however, problems brought about by skewed coordinate axes due to significant correlations among the variables are somewhat ameliorated by procedures, such as genetic algorithms, used for variable selection. While such procedures tend to remove highly correlated variables this may not always be the case so that coordinate system skew can still be a problem. However, if the variables are not too correlated the skew of a coordinate system will not significantly influence the overall results. A methodology is described in Subsection 3.3.1 that is based on molecular similarities and includes coordinate system non-orthogonality in a natural way.

2.3.3. Proximity Measures for Continuous-Valued Vectors

Because of the continuous nature of the vector components described in this section, other types of distance and similarity measures have been used. While the Hamming distance (*see* Eq. (20)) also applies for continuous vectors, Euclidean distances are usually used

$$\begin{aligned} d_{\text{Euc}}(\mathbf{v}_\text{A}, \mathbf{v}_\text{B}) &= \| \mathbf{v}_\text{A} - \mathbf{v}_\text{B} \| \\ &= \sqrt{\langle (\mathbf{v}_\text{A} - \mathbf{v}_\text{B}), (\mathbf{v}_\text{A} - \mathbf{v}_\text{B}) \rangle} \\ &= \sqrt{\sum_{k=1}^{n} \left(v_\text{A}(x_k) - v_\text{B}(x_k) \right)^2} \end{aligned} \tag{42}$$

In some instances, however, Minkowski distances are employed

$$d_{\text{Minkow}}(\mathbf{v}_\text{A}, \mathbf{v}_\text{B}) = \| \mathbf{v}_\text{A} - \mathbf{v}_\text{B} \|_{\ell_r} = \left[\sum_{k=1}^{n} \left| v_\text{A}(x_k) - v_\text{B}(x_k) \right|^r \right]^{\frac{1}{r}}, \tag{43}$$

where $r \geqslant 0$. Minkowski distances include both Hamming ($r = 1$) and Euclidean ($r = 2$) distances as special cases. Continuous distances can be converted into similarities using an appropriate monotonically decreasing function of distance, d, such as $\exp(-\eta \cdot d)$ or $1/(1 + \eta \cdot d)$, which both map to the unit interval, [0,1] for finite, non-negative values of η.

The most prevalent among the similarity coefficients is the so-called *cosine similarity index* or *correlation coefficient*. For the field-functions discussed in Subsection 2.4 it is usually called the *Carbó similarity index*

$$S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\sqrt{\|\mathbf{v}_A\|^2 \cdot \|\mathbf{v}_B\|^2}}, \qquad = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\|\mathbf{v}_A\| \cdot \|\mathbf{v}_B\|} \tag{44}$$

where the term in brackets in the numerator is the *inner product* of the two vectors

$$\langle \mathbf{v}_A, \mathbf{v}_B \rangle = \sum_{k=1}^{n} v_A(x_k) \cdot v_B(x_k) = \|\mathbf{v}_A\| \cdot \|\mathbf{v}_B\| \cos(\mathbf{v}_A, \mathbf{v}_B) \tag{45}$$

and their magnitudes are given by the Euclidean norm

$$\|\mathbf{v}_X\| = \sqrt{\langle \mathbf{v}_X, \mathbf{v}_X \rangle} = \sqrt{\sum_{k=1}^{n} v_X(x_k)^2}, \quad X = A, B. \tag{46}$$

It is important to note that the expressions in the latter two equations implicitly assume that the basis set used to describe the vectors is orthonormal.

As the similarity index is origin dependent there typically is a difference between the values computed for the cosine similarity index and correlation coefficients, since the latter is always computed at the mean of the of the data. Moreover, if the components of the vectors are all non-negative then $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B)$ is also non-negative. When this is not the case, $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B)$ may become negative, a situation that also obtains for the other similarity indices discussed in the remainder of this section. Maggiora et al. [53] have treated this case in great detail for continuous field functions, but the arguments can be carried through for finite vectors as well.

As has been pointed out numerous times, if $\mathbf{v}_A = \kappa \mathbf{v}_B$, then $S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) = 1$ for all κ. This prompted Hodgkin and Richards [60] to define a slightly modified form of molecular similarity, usually called the Hodgkin similarity index, that does not suffer from this problem, namely,

$$S_{\mathrm{Hod}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\frac{1}{2}\left(\|\mathbf{v}_A\|^2 + \|\mathbf{v}_B\|^2\right)}. \tag{47}$$

Petke [41] has developed two additional indices that bound both the Carbó and Hodgkin similarity indices, namely

$$S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\min\left(\|\mathbf{v}_A\|^2, \|\mathbf{v}_B\|^2\right)} \tag{48}$$

and

$$S_{\mathrm{Pet}_{\max}}(\mathbf{v}_A, \mathbf{v}_B) = \frac{\langle \mathbf{v}_A, \mathbf{v}_B \rangle}{\max\left(\|\mathbf{v}_A\|^2, \|\mathbf{v}_B\|^2\right)}, \tag{49}$$

that are analogous to those given, respectively, in Eqs. (28) and (29) for the case of sets or binary vectors. Recently, a comprehensive analysis has been given for continuous, field-based functions of all of the similarity coefficients of this general form, which characterizes their linear ordering and their upper and lower bounds [53] (*see* Eq. (66)). Their approach can be taken over in its entirety to the case of finite-dimensional vectors covered in this section. Thus, the bounds of the similarity indices in Eqs. (44), (47), (48), and (49), are given by

$$0 \leqslant S_{\mathrm{Pet}_{\max}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Hod}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Car}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B) \leqslant \infty. \tag{50}$$

All of the indices except $S_{\mathrm{Pet}_{\min}}(\mathbf{v}_A, \mathbf{v}_B)$ have upper bound of unity.

2.4. Field-Based Functions

Many methods exist for assessing 3-D molecular similarity. Good and Richards [61] and Lemmen and Lengauer [62] provide comprehensive reviews of most of the methods in use today, a large class of which utilizes some form of vector-based representation of 3-D molecular features such as 3-D pharmacophores [33, 63] and various types of 3-D shape descriptors [64]. The components of these vectors can be binary, integer, categorical, or continuous as discussed in the previous sections. Most 3-D methods, however, involve some type of direct alignment of the molecules being considered. Early on RMS deviations between specific atoms in the molecules being compared were employed, but this required identifying the key atoms, a non-trivial computational task. A variety of other 3-D methods exist [62], but the bulk of the 3-D methods utilize some form of field-based function to represent the fields or *pseudo*-fields surrounding the molecules that can be either continuous or discrete. Examples include "steric," electrostatic potential and "lipophilic" fields [39]. A novel, albeit lower resolution approach, based on ellipsoidal Gaussians has recently been developed [65] and shows great promise as a means for handling very large sets of molecules. Several workers have also developed a field-based methodology for directly aligning molecules based

upon their electric fields [41, 60] that differs from the usual scalar potential fields that are typically matched.

Interestingly, there is a close analogy between the alignment of 3-D molecular fields and the determination of maximum common substructures of two chemical graphs (*see* Subsection 2.1). Both cases involve the search for optimal overlays or alignments. The former requires continuous optimizations of non-linear similarity indices that give rise to large numbers of solutions and to great difficulties in clearly identifying the global maximum or "best" solution (*see* Subsection 2.4.3). The latter requires discrete optimizations, but the problem is NP complete and thus does not scale well computationally.

A major factor differentiating 3-D from 2-D similarity methods, regardless of the type of 3-D method employed is the need to account in some manner for conformational flexibility. There are two ways this is generally accomplished. One method involves carrying out a conformational analysis and selecting a subset of "appropriate" conformations for each molecule. All pairwise alignments of the selected conformations are then computed [37]. The other method involves some form of conformational search carried out simultaneously with the alignment process [66, 67]. Because of its importance in similarity-based alignments of molecules the remainder of the discussion in this section will focus on field-based methods.

2.4.1. Representation of Molecular Fields

Field-based methods generally utilize linear combinations of appropriate functions that are associated in some way with the atoms of the molecule under study:

$$F_{\mathrm{A}}^{\alpha}(\mathbf{r}) = \sum_{i \in \text{atoms}} a_i^{\alpha} f_i(\mathbf{r}), \tag{51}$$

where "α" designates the specific type of field or property being considered. The coefficients a_i^{α} weight the atom-based functions and in many cases are used to characterize specific properties attributed to the individual atoms (*vide infra*). Unnormalized, spherically symmetric Gaussian functions, "Gaussians" for short, are by far the most ubiquitous functions used in field-based applications:

$$f_i(\mathbf{r}) = \exp\left(-\kappa_i |\mathbf{r} - \mathbf{R}_i|^2\right), \tag{52}$$

where $\mathbf{R}_i$ is the location of the Gaussian, generally at an atomic center, and κ_i is its "width," which is the reciprocal of the variance, that is $\kappa_i = 1/\sigma_i^2$. The variance is sometimes referred the orbital radius, $\rho_i = \sigma_i^2$, of a Gaussian [68]. As $\kappa_i \to 0$, $f_i(\mathbf{r})$ becomes more spread out and conversely as $\kappa_i \to \infty$, $f_i(\mathbf{r})$ becomes sharper until, in the limit, it approaches an infinitely sharp delta function. In the latter case, atoms are essentially represented as points, while

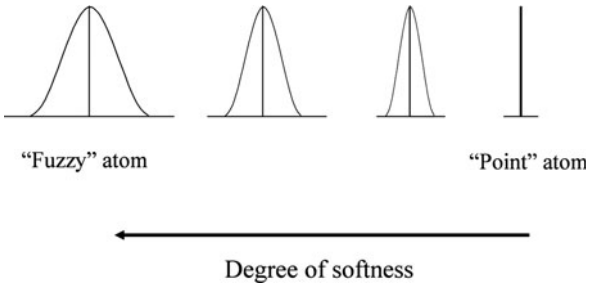


Fig. 6. Gaussian curves as a function of increasing width. As the degree of softness increases the curves represent "fuzzy" atoms and as the degree of softness deceases the Gaussian converges to a "point" atom model.

in the former case, they are represented as "soft spheres" as illustrated in Fig. 6. Although applications utilizing ellipsoidal Gaussians, which represent a generalization of spherically-symmetric Gaussians, are finding increasing application in chem-informatics [65], the focus of the current chapter will remain on the former more ubiquitous and simpler functions.

A useful property of Gaussians is that the integral of the product of two Gaussians [69] is given by another Gaussian that is a function of their distance of separation

$$\int f_i(\mathbf{r}) \cdot f_j(\mathbf{r}) d^3\mathbf{r} = \left(\frac{\pi}{\kappa_i + \kappa_j}\right)^{3/2} \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right). \tag{53}$$

Thus, the "overlap" of two molecules, A and B, with respect to the field of property α, $\Omega(F_A^\alpha, F_B^\alpha)$, is given by

$$\begin{aligned} \Omega(F_A^\alpha, F_B^\alpha) &= \int F_A^\alpha(\mathbf{r}) \cdot F_B^\alpha(\mathbf{r})\, d^3\mathbf{r} \\ &= \sum_{i \in A} \sum_{j \in B} a_i^\alpha \cdot b_j^\alpha \int f_i(\mathbf{r}) f_j(\mathbf{r})\, d^3\mathbf{r} \\ &= \sum_{i \in A} \sum_{j \in B} a_i^\alpha \cdot b_j^\alpha \left(\frac{\pi}{\kappa_i + \kappa_j}\right)^{3/2} \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right) \end{aligned}, \tag{54}$$

$$\Omega(F_A^\alpha, F_B^\alpha) = \sum_{i \in A} \sum_{j \in B} \tilde{a}_i^\alpha \cdot \tilde{b}_j^\alpha \exp\left(-\frac{\kappa_i \kappa_j}{\kappa_i + \kappa_j} \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right), \tag{55}$$

where the modified coefficients, $\tilde{a}_i^\alpha$ and $\tilde{b}_j^\alpha$, are obtained by including the square root term equally into the two field (i.e., property) coefficients a_i^α and b_j^α given in Eq. (54). In most cases the width parameters, κ_i and κ_j, are chosen to be the same for all atoms.

Equation (55) is a general form that is used in a number of field-based approaches to 3-D molecular alignment and similarity.

For example, in the program Seal [70] the coefficients given either in Eq. (54) or Eq. (55) are subsumed into a single "property coefficient," $\tilde{a}_i^{\alpha} \cdot \tilde{b}_j^{\alpha} \Rightarrow w_{i,j}$, which may account for the effect of multiple types of properties,

$$\Omega_{\text{'Seal'}}(\text{A}, \text{B}) = \sum_{i \in \text{A}} \sum_{j \in \text{B}} w_{i,j} \exp\left(-\kappa \left|\mathbf{R}_i - \mathbf{R}_j\right|^2\right). \tag{56}$$

The exponential coefficient, κ, determines the spread of the Gaussian and is taken be identical for all atom pairs. Some methods assign property values directly to the coefficients $\tilde{a}_i^{\alpha}$ and $\tilde{b}_j^{\alpha}$ [67, 71].

An alternative approach [37] treats the steric and electrostatic potential fields directly. The steric field is generally given by an expression similar to that in Eq. (51),

$$F_{\text{A}}^{\text{st}}(\mathbf{r}) = \sum_{i \in \text{atoms}} a_i^{\text{st}} f_i(\mathbf{r}), \tag{57}$$

where the coefficients are usually taken to be unity, that is $a_i^{\text{st}} = 1$, the field functions, $f_i(\mathbf{r})$, are usually taken to be Gaussians (*see* Eq. (52)), and the width parameters, κ_i, are either held constant for all atoms or are adjusted for each specific "atomic environment" [37]. In the case of the molecular electrostatic potential ("el") field

$$F_{\text{A}}^{\text{el}}(\mathbf{r}) = \sum_{i \in \text{atoms}} \frac{q_i}{|\mathbf{r} - \mathbf{R}_i|} \tag{58}$$

the $1/r$ term, which becomes singular at each atomic nucleus, presents a computational problem that was solved by Good et al. [72], who developed a Gaussian expansion of the "$1/r$" term,

$$\frac{1}{|\mathbf{r} - \mathbf{R}_i|} \approx \sum_{k \in \text{A}_i} c_k f_{i,k}(\mathbf{r}), \tag{59}$$

that significantly expedites computations. In this expression $f_{i,k}(\mathbf{r})$ is the k-th Gaussian in the expansion of $1/r$ about the i-th atom, A_i, of molecule A. The expansion usually consists of two or three terms, and the expansion coefficients, c_k, are obtained by least-squares minimization. Note that the width parameters, κ_k, are independent of the atom center and differ significantly from each other in order to fit the $1/r$ term with sufficient accuracy [72]. Substituting Eq. (59) into Eq. (58) converts it into a sum of Gaussians, and thus most field-based similarity measures (*vide infra*) only require calculation of Gaussian overlap integrals (*see* e.g., Eqs. (53) and (54)) when dealing with steric or electrostatic potential fields. Thus, many of the issues that plague similarity calculations carried out within a discrete lattice framework are no longer a problem in the case of continuous field-based functions.

2.4.2. Field-Based Similarity Indices

Field-based similarities are usually evaluated by the cosine or correlation function similarity measure employed initially by Carbó and Calabuig [73] to compute molecular similarities based upon quantum mechanical wavefunctions. Such a measure, which is usually called a Carbó similarity index, is given by

$$S_{\mathrm{Car}}\left(F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\left\|F_{\mathrm{A}}^{\alpha}\right\| \cdot \left\|F_{\mathrm{B}}^{\alpha}\right\|} = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\sqrt{\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2} \cdot \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}}}, \tag{60}$$

where the inner product in the numerator is now given by an integral rather than a summation since the objects considered here are field functions, F_{A}^{α} and F_{B}^{α}, not vectors, although strictly speaking functions are equivalent to infinite dimensional vectors,

$$\left\langle F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right\rangle = \int F_{\mathrm{A}}^{\alpha}(\mathbf{r}) \cdot F_{\mathrm{B}}^{\alpha}(\mathbf{r}) d^{3}\mathbf{r}, \tag{61}$$

and the Euclidean norm of the functions is given by

$$\left\|F_{X}^{\alpha}\right\| = \sqrt{\int F_{X}^{\alpha}(\mathbf{r})^{2} d^{3}\mathbf{r}}, \quad X{=}A, B. \tag{62}$$

Note the similarity of Eqs. (45) and (46) to Eqs. (61) and (62). The main difference between them arises in the way in which the inner products are evaluated. Also, as was the case for vectors if the field functions are non-negative functions $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha})$ will be non-negative. When this is not the case, however, $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha})$ may become negative, a situation that also obtains for the other similarity indices discussed in the remainder of this section. Maggiora et al. [53], have treated this case in great detail for continuous field functions, but the arguments can be carried through for finite vectors as well (*vide supra*).

As discussed in the previous section for vectors, if F_{A}^{α} and F_{B}^{α} differ only by a constant, that is if $F_{\mathrm{A}}^{\alpha} = K \cdot F_{\mathrm{B}}^{\alpha}$, then $S_{\mathrm{Car}}(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}) = 1$ regardless of the specific form of the functions. While this is not a likely occurrence in practical applications, Hodgkin and Richards [60] nonetheless defined a slightly altered similarity measure, usually referred to as the Hodgkin similarity index, which is not affected by this problem and is given by

$$S_{\mathrm{Hod}}\left(F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\, F_{\mathrm{B}}^{\alpha}\right\rangle}{\frac{1}{2}\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2} + \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)}, \tag{63}$$

where the terms in the denominator of Eqs. (60) and (63) are, respectively, the geometric and arithmetic means of the squared norms of F_{A}^{α} and F_{B}^{α}. As has been shown by Maggiora et al. [53],

a family of similarity indices can be defined in terms of the means of the squared norms in their denominators.

The Petke indices, defined earlier for vectors (*see* Eqs. (48) and (49)), are given by

$$S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right\rangle}{\min\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2}, \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)} \tag{64}$$

and

$$\mathrm{S}_{\mathrm{Pet}_{\max}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right) = \frac{\left\langle F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right\rangle}{\max\left(\left\|F_{\mathrm{A}}^{\alpha}\right\|^{2},\ \left\|F_{\mathrm{B}}^{\alpha}\right\|^{2}\right)}. \tag{65}$$

All of the same bounding properties described in the previous section for vectors (*see* Eq. (50)) obtain here as well, including the fact that all of the indices except $S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha},\ F_{\mathrm{B}}^{\alpha}\right)$ are bounded from above by unity [53]:

$$0 \leqslant S_{\mathrm{Pet}_{\max}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Hod}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Car}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant S_{\mathrm{Pet}_{\min}}\left(F_{\mathrm{A}}^{\alpha}, F_{\mathrm{B}}^{\alpha}\right) \leqslant \infty. \tag{66}$$

None of the cosine/correlation-like similarity indices or their complements (*see* Subsection 2.5 on Dissimilarity Measures) are true metrics; that is, they do not obey the distance axioms. Petitjean [74, 75], however, has developed a distance-based methodology, but it has not been applied in many cases.

Similarity indices corresponding to different fields can be combined into an overall similarity index, for example

$$S(F_{\mathrm{A}}, F_{\mathrm{B}}) = \lambda \cdot S(F_{\mathrm{A}}^{\mathrm{st}}, F_{\mathrm{B}}^{\mathrm{st}}) + (1 - \lambda) \cdot S(F_{\mathrm{A}}^{\mathrm{el}}, F_{\mathrm{B}}^{\mathrm{el}}), \tag{67}$$

where S is the Carbó, Hodgkin, Petke, or other appropriate index [53] and λ is the weighting coefficient. Mestres et al. [37], have used a value ($\lambda \approx 0.66$), arrived at pragmatically, that weights steric to electrostatic-potential similarity in a 2:1 ratio.

2.4.3. Deriving Consistent Multi-molecule Alignments

As has been shown by Mestres et al. [29], the optimal solution, $S_X(F_{\mathrm{A}}, F_{\mathrm{B}})_1$, may not correspond to the correct "experimentally-derived" molecular alignment. To address this problem, these authors developed the concept of *pairwise consistency*, which is depicted in Fig. 7. Consider the similarities of three molecules A, B, and C. Suppose molecule A is the reference molecule, which is held fixed, and molecules B and C are the adapting molecules. Now determine the optimal similarity solutions for $S(F_{\mathrm{A}}, F_{\mathrm{B}})_1$ and $S(F_{\mathrm{A}}, F_{\mathrm{C}})_1$ using an appropriate similarity index. Both molecules B and C are now aligned to molecule A. Keeping their positions relative to molecule A fixed, compute $S(F_{\mathrm{B}}, F_{\mathrm{C}})^{*}$, which is not necessarily equal to the optimized solution, that is $S(F_{\mathrm{B}}, F_{\mathrm{C}})^{*} \neq S(F_{\mathrm{B}}, F_{\mathrm{C}})_1$. Pairwise consistency holds only in the

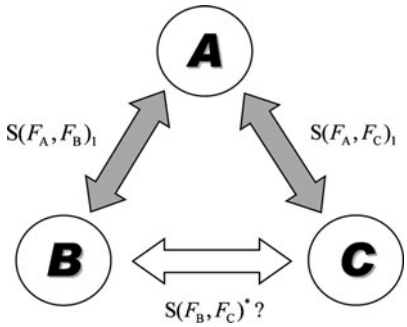


Fig. 7. Depiction of *pairwise consistency* among three molecules, A, B and C.

case when *equality* obtains, otherwise the solutions are said to be pairwise inconsistent. Sometimes pairwise consistency is obtained when one of the lower similarity solutions is considered, say for example $S(F_A, F_C)_2$. In such cases, the alignments given by $S(F_A, F_B)_1$,$S(F_A, F_B)_2$, and $S(F_B, F_C)^* = S(F_B, F_C)_1$ are assumed to be the correct alignments. In many cases, it is not possible to identify pairwise consistent set of solutions. In such cases, the fields of three molecules are *simultaneously aligned* using the average of the pairwise similarities

$$S(F_A, F_B, F_C) = \tfrac{1}{3}[S(F_A, F_B) + S(F_A, F_C) + S(F_B, F_C)]. \qquad (68)$$

This procedure automatically generates pairwise consistent solutions, and it can be continued to higher orders until consistent solutions are obtained to all orders, a computationally very demanding task that has not been pursued in most cases since the ternary similarities are generally sufficient for molecular design purposes.

Note that any 3-D similarity method that involves molecular superpositioning can be used to assess the consistency of multi-molecule alignments. However, 3-D methods that do not involve superpositions (e.g., see the interesting recent work by Ballester and Richards [76]) are not suitable for this type of analysis.

2.4.4. Addressing Conformational Flexibility

As noted in the introduction of this section, computing 3-D molecular similarities of a set of molecules requires physically aligning the appropriate fields of the molecules, while accounting for their conformational flexibility, which can be accomplished in two ways either by rigid body superpositions of selected conformations of each of the molecules being aligned or by simultaneous conformational sampling during the alignment process. In the rigid-body case, one molecule is generally chosen as the *Reference Molecule*, which remains fixed, while the other *Adapting Molecule* is translated and rotated until a maximum of the similarity index is obtained. Since the similarity index is a non-linear function it generally has multiple solutions,

$$S(F_A, F_B)_1 \geqslant S(F_A, F_B)_2 \geqslant \cdots \geqslant S(F_A, F_B)_k \geqslant \cdots \tag{69}$$

although it is difficult to know if the global maximum has been attained. To increase the chances that all of the best solutions are obtained, multiple starting geometries are usually sampled.

An added difficulty in rigid-body alignment is that all "relevant" conformations, N, of the reference molecule must be aligned with all "relevant" conformations, M, of the adapting molecule – $N \times M$ alignments must be carried out where, as discussed above, each alignment involves multiple starting geometries. This is a significant computational burden for the alignment of a single pair of molecules, and thus carrying out alignments for a large set of molecules is not computationally feasible at this time.

There are some approaches that hold promise for speeding up the computations. A novel procedure based upon Fourier transforms was developed by Nissink et al. [77], and used by Lemmen et al. [71] The method separates the translational and rotational motions needed to align pairs of molecules and thus allows the separate optimization of each, thereby facilitating the overall alignment process. While this certainly speeds up the computations, it does not significantly alter the significant time requirements of rigid body alignments.

An alternative approach that combines conformational searching with similarity-based structure alignment perhaps holds more promise in terms of speeding up the process of aligning conformationally flexible molecules. In contrast to the rigid-body alignment process described above, in this case both molecules are treated on an equal footing and are allowed to move and conformationally flex. In the approach of Blinn et al. [66], which is similar to that developed by Labute [67], the energy of the combined system of the two molecules being aligned, $E_{A,B}^{total}$, is given by

$$E_{A,B}^{total} = E_A^{conf} + E_B^{conf} + E_{A,B}^{sim}, \tag{70}$$

where E_A^{conf} is the conformational energy of molecule A, E_B^{conf} is the conformational energy of molecule B, and $E_{A,B}^{sim}$ is a pseudo-energy penalty term, which is given by

$$\begin{aligned} E_{A,B}^{sim} &= K_{sim} \cdot [1 - S(F_A, F_B)] \\ &= K_{sim} \cdot D(F_A, F_B) \end{aligned}, \tag{71}$$

where K_{sim} is an adjustable proportionality constant, which lies in the range of 5–20 kcal/mol. The dissimilarity $D(F_A, F_B)$ (*see* Subsection 2.5) is used rather than similarity since the penalty term should vanish when the fields of the two molecules are in perfect alignment that is when $S(F_A, F_B) = 1 \rightarrow D(F_A, F_B) = 0 \rightarrow E_{A,B}^{sim} = 0$. Alternatively, the maximum penalty should be assessed when

$$S(F_A, F_B) = 0 \rightarrow D(F_A, F_B) = 1 \rightarrow E_{A,B}^{sim} = K_{sim}. \tag{72}$$

Other forms for the pseudo-energy penalty term have also been investigated [66, 67]. In any case, the pseudo-energy penalty term acts as a constraint on the overall energy of the system, which is a balance between favorable conformational energies and overall molecular alignment as measured by field-based similarity (dissimilarity).

While the approaches noted above use some type of stochastic procedure to explore the similarity-constrained conformational space [66, 67], they are not in principle restricted to such search methods. Molecular dynamics-based procedures are also possible, although to our knowledge none have as yet been developed to address this problem.

2.4.5. Multi-conformer-Dependent Similarities

It may be argued that 3-D similarities most closely represent the concept of molecular similarity. Since many molecules can attain multiple conformational states, it is not unreasonable to assume that these low-lying conformational states should play a role in molecular similarity. To date, 3-D similarity methods typically choose a single conformer, usually the one with the lowest conformational energy. This section describes a possible approach to the question "Given that a set of conformer-dependent similarities can be determined, how does one assign an aggregate similarity to the set of values"? It is not meant to be a finished work, but rather is meant to encourage further work and discussion on what undoubtedly will become a more important issue as computational methods improve and become faster (*see* e.g. the work described in [76]).

One, but certainly not the only, approach is to weigh the similarities by the joint probability of the two conformers,

$$\langle S(F_{\mathrm{A}}, F_{\mathrm{B}})\rangle = \sum_{i=1}^{n_{\mathrm{A}}} \sum_{j=1}^{n_{\mathrm{B}}} \mathrm{Prob}\left(\mathrm{A}_i, \mathrm{B}_j\right) \cdot \mathrm{S}(F_{\mathrm{A}_i}, F_{\mathrm{B}_j}), \tag{73}$$

where A_i is the i-th conformational state of molecule A and B_j the j-th conformational state of molecule B. In the case where the conformations of each of the two molecules are determined independently, as described in Subsection 2.4.4,

$$\mathrm{Prob}\left(\mathrm{A}_i\,, \mathrm{B}_j\right) = \mathrm{Prob}(\mathrm{A}_i) \cdot \mathrm{Prob}\left(\mathrm{B}_j\right). \tag{74}$$

The two conformational-state probabilities can be estimated using a Boltzmann or other suitable type of probability function; the former is given by [78]

$$\mathrm{Prob}(\mathrm{M}_\ell) = \frac{\exp(-\kappa \cdot \Delta E_{\mathrm{M}_\ell})}{Z_{\mathrm{M}}}, \quad \mathrm{M} = \mathrm{A},\ \mathrm{B},\ \mathrm{C}, \ldots; \quad \ell = 1, 2, \ldots, n_{\mathrm{M}} \tag{75}$$

where κ is Boltzmann's constant, $\Delta E_{\mathrm{M}_\ell}$ is the energy (in kcal/mol) of the i-th conformation of M *relative* to its lowest energy

conformation, and Z_M is the conformational partition function for molecule M, i.e.

$$Z_M = \sum_{\ell=1}^{n_M} \exp(-\kappa \cdot \Delta E_{M_\ell}), \quad M = A, B, C, \ldots \tag{76}$$

where n_M is the number of conformational states considered for molecule M. Since n_M is usually less than the actual number of conformational states, Z_M is only an estimate of the true conformational partition function. However, this may not be the most serious limitation of the proposed approach as conformational energies may also be poorly estimated and solvent effects, when they are explicitly considered, also can represent a serious source of error (*vide infra*).

For entirely practical reasons of computational efficiency, conformational energies are usually computed in the gas phase using molecular mechanics potential-energy functions [79]. However, recent advances in computational methods have made solvent-based conformational energetics computationally feasible [80]. Methods have also been developed for computing Gibbs free energies [81], which are physically more realistic. However, even in cases where conformational energetics are computed with reasonable accuracy, the number of conformers considered is usually only a subset of the possible conformers such as those, for example, based on clustering (*see* discussion in Subsection 2.2.4). Thus, the calculated probabilities are highly approximate. Nevertheless, this approach accounts in some, albeit very approximate, fashion for the role that multiple conformations can play in MSA. Whether this more elaborate approach produces better results than those obtained using single low-energy conformations has yet to be investigated.

Considering the combinatoric explosion that can arise in the "independent conformer" approach, it may be more effective to compute conformationally-weighted similarities in a more direct manner. Such an approach may be developed from that described in Subsection 2.2.4 (*see* Eqs. (70)–(72)) [66] or some other variant [67] that combines conformational searching simultaneously with similarity-based alignment. Although, conformational independence (*vide supra*) is lost in this approach, it may not be an impediment to obtaining computationally-feasible, conformationally-weighted similarities.

A related but "softer" approach (*see* e.g., Petit et al. *submitted* [82], for a discussion of soft approaches to the Rule of Five that is relevant here) is to consider the conformationally-dependent distribution of similarity values between each pair of molecules under consideration in a given study, and then to compare the resulting similarity or cumulative similarity distributions [83] directly or with respect to several relevant distributional parameters such

as the mean, median, standard deviation, or other statistical moments [84].

2.5. Dissimilarity Measures

Dissimilarity is generally taken to be the complement of similarity, that is

$$D(\mathrm{A},\mathrm{B}) = 1 - S(\mathrm{A},\mathrm{B}). \tag{77}$$

While this is mathematically reasonable, and is thus used extensively, psychologically the two concepts are not so simply related. This stems from the fact that assessing the similarity of two objects is easier than assessing their dissimilarity. As two objects become less and less similar a point is reached, say for a similarity value of 0.35, below which it is very difficult to assign a value to their similarity. Since dissimilarity is just the complement of similarity it follows that humans can only properly assess the dissimilarity of two objects if they are not too dissimilar. Thus, even though we can assign dissimilarities ~1.0 using Eq. (77) its "meaning" is not easily grasped. This is why we have focused our discussion on similarity rather than dissimilarity, even though the concept of dissimilarity has important practical applications in studies of molecular diversity. Martin [85] has edited an interesting account of the development and implementation of the concepts of molecular dissimilarity and diversity in cheminformatics and combinatorial chemistry. Seilo [86] has also investigated some of the issues associated with comparing dissimilar molecules.

3. Chemical Spaces

This section on chemical spaces, while important, is not presented with the same level of mathematical detail as given in earlier sections. The object here is to provide a general discussion of some of the important characteristics of these spaces. A recent review by Medina-Franco et al. [87] presents an excellent overview of many aspects of chemical spaces relevant to drug design research; Oprea and Gottfries [88] published the first paper on navigating chemical spaces.

The concept of a chemical space derives from the notion of a space used in mathematics and is taken here to be a set of molecules along with one or more relationships defined on the elements (i.e. molecules) of the set. The nature of a given chemical space depends, directly or indirectly, on how the molecular information is represented (Subsection 2); the representation used strongly influences what can be known about the set of molecules under study. Figure 8 illustrates the dramatic effect that different molecular representations can have on the distribution of compounds in chemical space (*See* figure legend for details and

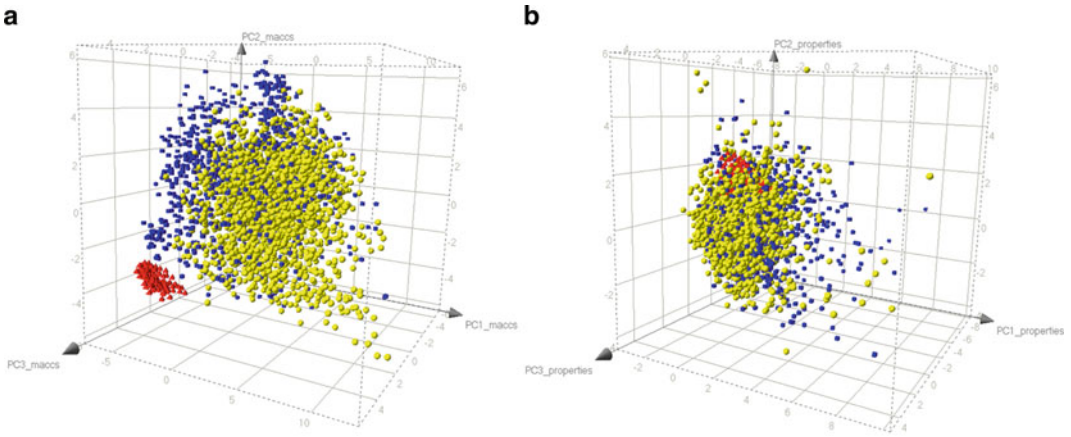


Fig. 8. Comparison of the 125 new molecules (*red triangles*), 1,490 approved drugs (*blue squares*) and a representative set of 2,000 diverse compounds (*yellow spheres*) from Molecular Libraries Small Molecules Repository (MLSMR) (*Original figure provided courtesy of Dr. José Medina-Franco*). (**a**) Depiction of a visual representation of the chemical space obtained by PCA of the similarity matrix computed using MACCS keys and Tanimoto similarity. The first three PCs account for 62.1% of the variance. (**b**) Depiction of a visual representation of the property space obtained by PCA of six scaled physicochemical properties (MW, RB, HBA, HBD, TPSA, and SlogP). The first three PCs account for 84.3% of the variance.

Subsections 3.1 and 3.1.1 for a discussion of how chemical spaces are depicted and how reduced-dimensional chemical spaces are constructed.). Panel (a) of the figure shows that the set of 125 new molecules (red triangles) is *structurally different* from the 1,490 approved drugs (blue cubes) [89] and the representative set of 2,000 compounds from the MLSMR [90]. Panel (b) indicates that the new library is located within a region of the physicochemical property-based chemical space that is associated with bioactive molecules. Both of these pieces of information are useful. Nevertheless, the significant differences in the distribution of compounds in these two spaces clearly show the crucial role that representation plays in defining chemical spaces.

Importantly, unlike the case in physics, the underlying relationships in chemical spaces are not invariant to representation. For example, neighborhood relationships that obtain in one chemical space may not also obtain in another chemical space [2] (*Cf.* Patterson et al. [91]). This is shown in Fig. 9, which schematically depicts the chemical spaces generated for the same set of molecules using two different representations (♣ and ♠). As is seen in the figure, molecules that are nearest-neighbors with respect to one representation may not even be close neighbors in the other (*See*, for example, mappings ❶ and ❷ and mappings ❷ and ❸ of Fig. 9). Thus, there is *loss of topological invariance*, which is a much more serious condition than the loss of purely geometric features such as the distances between molecules or the angles between vectors representing the locations of molecules in

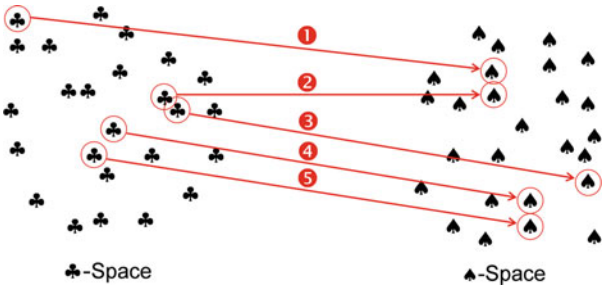


Fig. 9. An example illustrating that chemical spaces are representation dependent. Five point-to-point mappings are indicated on the figure. Mappings ❶ and ❷ show that two molecules that are well separated in ♣-Space may be nearest-neighbors in ♠-Space, while mappings ❷ and ❸ show that two molecules that are nearest-neighbors in ♣-Space are well separated in ♠-Space. Mappings ❹ and ❺ show that neighborhood relationships may also be maintained by mappings between the two spaces.

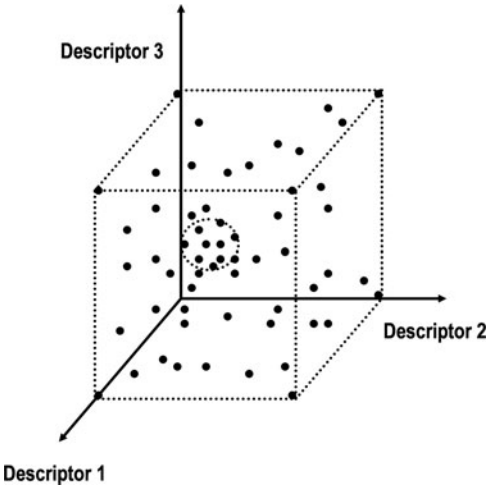


Fig. 10. In coordinate-based chemistry space, points in close proximity are considered to be similar. For instance, the compounds within the sphere shown here are quite similar to each other compared to compounds in the extremities of this 3-dimensional chemistry space.

chemical spaces. Loss of topological invariance can have dire consequences in subset selection procedures since it can change the rank ordering of neighboring compounds with respect to chemical spaces constructed using different similarity measures [2].

3.1. Dimensionality of Chemical Spaces

Chemical spaces can be grouped into two broad classes, namely, *coordinate-based* and *coordinate-free*. In coordinate-based chemical spaces molecules are represented as points distributed throughout the space as illustrated in Fig. 10. Points in close proximity are considered to represent similar molecules,

while distant points represent dissimilar molecules. An important feature of coordinate-based chemical spaces is that the *absolute position* of a molecule within the space is known, not just its position relative to the other molecules in the space. This is not the case with coordinate-free chemical spaces. In such spaces the relationship of a given molecule to its near and far neighbors is known but not its location within the space. Thus, finding "compound voids" in a coordinate-free chemical space is a much more difficult task than it is in a coordinate-based chemical space. An additional useful feature of coordinate-based chemical spaces is their ability to portray the distribution of compounds in ways that, in many cases, can enhance our understanding of the space. However, as is discussed in the following paragraph, the high-dimensionality of these spaces can frustrate attempts to visualize them.

The dimension of a coordinate-based chemical space is simply the number of independent variables used to define the space. As seen in earlier discussions, the dimension of such spaces can be quite large, and there are a significant number of examples where the dimension can exceed one million [33, 48]. Even for spaces of much lower dimension, say around ten or greater, the effects of the "*Curse of Dimensionality*" [92, 93] can be felt. Bishop [94] provides an excellent example, which shows that the ratio of the volume of a hypersphere inscribed in a unit hypercube of the same dimension goes to zero as the dimensionality goes to infinity. Actually, even for the ten-dimensional case the volume of the hypersphere is less than ten percent of that of the corresponding hypercube. Raghavendra and Maggiora [95] provide a more detailed discussion of some of the idiosyncratic behaviors of high-dimensional spaces [96]. In addition, most of the spaces of extremely high dimension are discrete since the number of "points" (i.e. molecules) in the space is finite, a feature that can present additional problems.

Although, it is possible in coordinate-free chemical spaces to rigorously construct coordinates within a Euclidean space for any set of molecules in the space, faithfully representing the intermolecular proximities may require that the space be of quite high-dimension, in an extreme case possibly equal to one less than the number of molecules in the set. As will be discussed in the following section, a number of methods exist for constructing low-dimensional Euclidean spaces for both high-dimensional coordinate-based and coordinate-free representations of molecules.

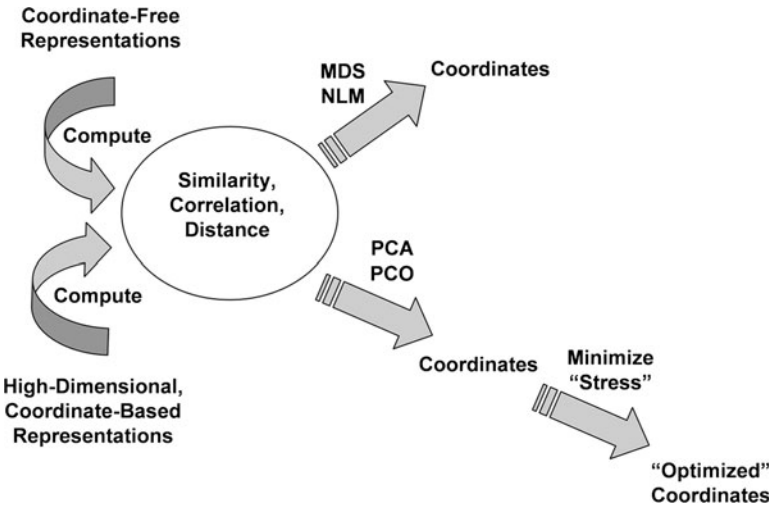
3.2. Constructing Reduced-Dimension Chemical Spaces [97]

This section provides a brief account on the construction of reduced-dimension chemical spaces for sets of molecules described by coordinate-free or by high-dimensional coordinate-based representations. Inherently low-dimensional chemical spaces such as those generated, for example, by BCUT descriptors

are not considered; the paper by Pearlman and Smith [35] should be consulted for a discussion of these descriptors. It is important to note that all of the issues surrounding the inherent non-orthogonality of coordinate systems described in Subsection 2.3.2 are applicable here as well, and that section should be consulted for further details.

Scheme 1 illustrates the various procedures for the construction of reduced-dimension coordinate systems – similarity (dissimilarity), correlation, or distance play a central role in these coordinate systems. The first step in reducing the dimension of either coordinate-free or high-dimensional coordinate-based representations is computation of some proximity measure of the similarity, correlation, or distance between all the pairs of molecules in the set of interest. This can be accomplished using the methods described earlier in this work. For example, in the coordinate-free case similarity can be computed using the graph-theoretical procedures described in Subsection 2.1, the field-based approach described in Subsection 2.4, or other less well-known approaches such as shape-group [98] and feature-tree [99] methods. In high-dimensional coordinate-based cases all of the vector-based approaches described in Subsection 2 are applicable (*N.B.* that field-based approaches can also fit under this rubric, since field-based functions can be considered to be infinite-dimensional vectors).

Once a proximity measure has been computed for all of the molecules, basically two paths exist for determining a lower-dimensional coordinate-based representation. In the upper path in Scheme 1 coordinates are determined using either multi-dimensional scaling (MDS) [11] or non-linear mapping



Scheme 1

(NLM) [13] procedures, both of which require minimization of some sort of error function. In the past, both procedures were somewhat limited and could only deal effectively with datasets of less than ~2000 molecules. In addition, they encountered difficulty in treating new sets of compounds that were not included in the original set without redoing the calculations for the entire augmented set. These limitations have been removed by the work of Agrafiotis and his colleagues [100, 101] who developed a clever neural-net approach that learned the non-linear mapping based upon the use of training sets of relatively small sample size (~1,000 compounds). Once the mapping function is learned new compounds can be mapped with relative ease.

The lower path is somewhat more complicated. The first step in the path involves either PCA [12] or principal-coordinate analysis (PCO) [12]. This step can be followed by optimization of a function that minimizes the error between the proximity measure computed in the reduced-dimension and full coordinate systems if desired. Xie et al. [102], recently published an interesting paper along these lines. Kruscal stress [103] is a widely used function in this regard, namely,

$$K_{\text{stress}} = \sqrt{\frac{\sum_i \sum_{j>i} \left(\hat{d}_{i,j} - d_{i,j}\right)^2}{\sum_i \sum_{j>i} \hat{d}_{i,j}^{\,2}}}, \tag{78}$$

where $\hat{d}_{i,j}$ is the distance computed in the reduced-dimension space and $d_{i,j}$ is the distance computed in the full space.

PCA is designed to deal directly with correlation matrices, but not directly with similarity or distance matrices. However, as pointed out by Kruscal [103], the similarity matrix (or other proximity matrix) can be treated as a normal data matrix upon which principal component analysis is performed, that is

$$\mathbf{S} \underset{\substack{\text{mean}\\ \text{center}\\ \text{columns}}}{\Rightarrow\Rightarrow} \bar{\mathbf{S}} \underset{\substack{\text{form}\\ \text{matrix}\\ \text{product}}}{\Rightarrow\Rightarrow} \bar{\mathbf{S}}^{\mathrm{T}}\bar{\mathbf{S}} \underset{\substack{\text{diagonalize}\\ \text{matrix}}}{\Rightarrow\Rightarrow} \mathbf{V}^{\mathrm{T}}\left(\bar{\mathbf{S}}^{\mathrm{T}}\bar{\mathbf{S}}\right)\mathbf{V} = \Lambda, \tag{79}$$

where the columns of the eigenvector matrix $\mathbf{V}$ are the principal components and the elements of the diagonal matrix Λ are the corresponding eigenvalues. The coordinates in the transformed PC coordinate system, usually called the "scores," are given by the matrix $\mathbf{T}$, where

$$\mathbf{T} = \bar{\mathbf{S}}\,\mathbf{V}. \tag{80}$$

Principal coordinate analysis [12] works in an analogous fashion except that the similarity matrix is used directly without the additional multiplications given in Eq. (79). Gower has described

the relationship between PCA and PCO [104]. Because both approaches utilize matrix diagonalization procedures, the size systems that they can practically treat are limited to ~2,000 molecules. This computational obstacle can be overcome for PCA using one of the neural net methods for determining principal components [105]. Benigni [106] described an analogous method based upon a matrix of Euclidean distances computed from high-dimensional vectors representing a set of molecules. Analogous dissimilarity-based methods have also been developed.

An important question is whether the proximity measures are compatible with those of these references addresses the important issue of whether the proximity measure is compatible with embedding in a Euclidean space. For example, satisfying the distance axioms do not guarantee that any distance matrix associated with a given set of molecules will be compatible, as the distance axioms are still satisfied in non-Euclidean spaces. Gower has written extensively on this important issue, and his work should be consulted for details [107–109]. Benigni [110] and Carbó [73] have also contributed interesting approaches in this area.

More recent work by a number of authors has further addressed the issue of embedding of high-dimensional data into lower dimensional spaces. These methods include the similarity-based abstract vector-space approach of Raghavendra and Maggiora [95], isometric mapping [111], local linear embedding [112], exploratory projection-pursuit [113], stochastic proximity embedding [114, 115], and eigenvalue-based methods [116].

3.3. Activity Cliffs and the Topography of Activity Landscapes

Chemical spaces and the activities of molecules in these spaces induce *activity landscapes*. In three dimensions, activity landscapes can be visualized as surfaces with features that are analogous to the Earth's topographical features – mountains, canyons, hills, valleys, cliffs, ridges, spires, plains, etc. Neglecting the Earth's curvature, which for small surface regions can be considered essentially flat, topographies can be represented by two rectilinear position coordinates and one rectilinear altitude coordinate. In activity landscapes, the two position coordinates give the location of a molecule within a two-dimensional chemical space and the altitude coordinate corresponds to the molecule's activity value with respect to a given biological or pharmacological assay. Thus, a number of different activity landscapes exist for a set of molecules in a given chemical space, one for each assay. Because many assays are of relatively low resolution, their activity landscapes will typically be of low resolution. Moreover, since chemical spaces are not invariant to representation, the nature of their associated activity landscapes will also be affected by the representation used to define the chemical space (*Cf.* Fig. 8). Lastly, chemical spaces are typically greater than two dimensions so the geographical analogy associated with the Earth, while familiar, is definitely a significant

simplification of the situation encountered in chemical spaces. Nevertheless, the information provided by simple 3-D models of activity landscapes still provides many useful insights that facilitate our understanding of the actual multi-dimensional case.

Until recently, it was generally assumed that small changes in activity are typically associated with small changes in molecular similarity [4]. In such cases, activity landscapes tend to resemble the rolling hills of Kansas. However, a growing number of studies have shown that this is not generally the case. In fact, activity landscapes appear to contain regions with "cliffs" that resemble the rugged landscape of Bryce Canyon more than that of Kansas [117]. Such "activity cliffs" arise when small changes in molecular similarity result in correspondingly large changes in activity. A recent editorial by Maggiora [118] suggests that activity cliffs play a significant role in the determination of quantitative structure–activity relations (QSAR). This editorial was followed up by several papers that provided additional discussion of this topic [119–121]. *Importantly, activity cliffs pinpoint regions of activity landscapes that contain maximum information on SARs.* This is so because small changes in molecular similarity make it easier to identify what features may be responsible for the dramatic shifts observed in activity. In contrast, it is difficult to ascertain just what features may be responsible for large activity shifts observed between two molecules that are very dissimilar. Figure 11, taken from the recent review by Bajorath et al. [122], provides a 3-D example based on data obtained from a set of Cyclooxygenase-2 inhibitors that illustrates the topography of a typical activity landscape and indicates the relationship between an activity cliff and its associated SAR. Note that the coordinates in the two-dimensional chemical space depicted in the figure are obtained by projection from a higher-dimensional chemical space.

3.3.1. Development of Structure–Activity Similarity Maps

Although the 3-D activity landscape portrayed in Fig. 11 has great intuitive appeal, it does not present an accurate picture of the true activity landscape, which is of much higher dimension. What is needed is a relatively simple representation of the data that can be visualized and analyzed and that captures a significant portion of the information contained in the activity landscape. Early work in this area by Shanmugasundaram and Maggiora [123] introduced the concept of a structure–activity–similarity map or *SAS map*. Figure 12 provides an example of such a map. The ordinate represents the "activity similarity," which is defined by

$$S_{\mathrm{Act}}(\mathrm{A},\mathrm{B}) = 1 - \frac{|\mathrm{Act}(\mathrm{A}) - \mathrm{Act}(\mathrm{B})|}{\mathrm{Act}_{\max} - \mathrm{Act}_{\min}}, \tag{81}$$

where Act(X) is the activity of compound A or B, typically given in terms of their pI_{50} values, and $\mathrm{Act}_{\max} - \mathrm{Act}_{\min}$ is the difference between the maximum and minimum activity values of the

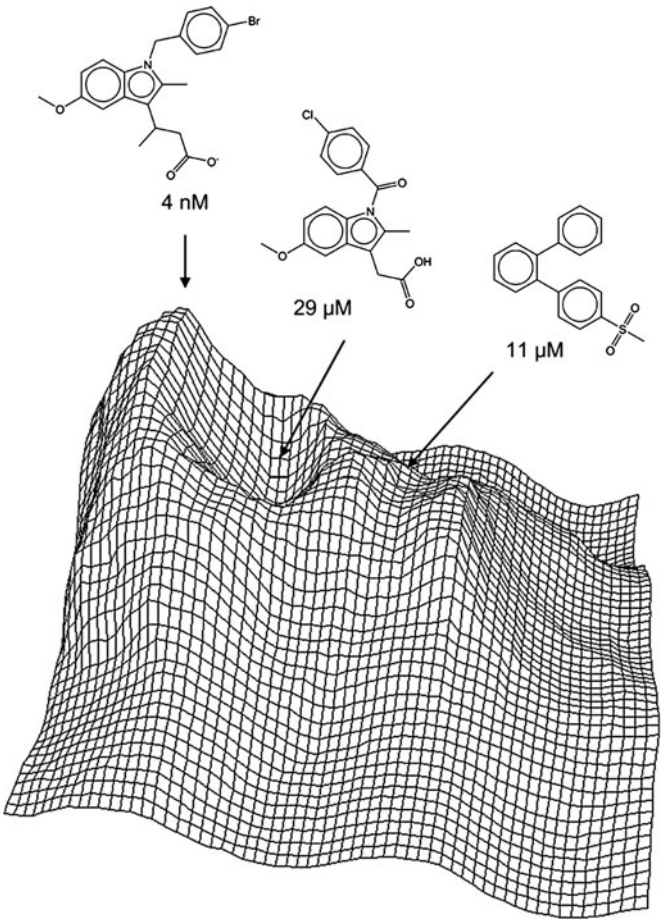


Fig. 11. Example of an activity landscape generated using data obtained from a set of cyclooxygenase-2 inhibitors (*Original figure provided courtesy of Prof. Jurgen Bajorath, see* ref. [122]). Note that the reference chemical space depicted here is a projection onto two dimensions of the actual higher-dimensional chemical space.

compounds in the dataset. This represents a "normalized" activity difference, although non-normalized activity differences can also be used in SAS maps. The abscissa represents the familiar "structure similarity"; any of the similarity measures discussed in this chapter can be used here.

Each of the 1,275 datapoints shown in Panel (a) of the figure represents a pairwise comparison with respect to the activity and structure similarity of each of the molecules in a small, prototypical dataset containing 51 molecules (*N.B.* that the number of distinct pairs obtained from N molecules is $N(N-1)/2$). The data points are color-coded by the activity of the most active compound of a given pair. Red circles denote pairs where at least one compound is active; yellow circles indicate pairs where at least one compound is moderately active; blue circles indicate pairs

where both compounds are inactive (or have low activity). Note that the scale for "Structure Similarity" does not run from zero to unity. This is because of the limited size and diversity of the dataset, since the minimum similarity between any pair of compounds is 0.58.

The SAS map is divided into four quadrants as shown in Panel (b) of the figure. Points located in the upper right quadrant (I) of the diagram correspond to pairs of compounds with both high activity and high structure similarity. The topography in this quadrant is relatively smooth, gently rolling hills. Compounds in this region behave in a "traditional" SAR fashion and reliable QSARs can generally be obtained from such compounds. Points in the lower right quadrant (II) correspond to pairs of compounds with high structure but low activity similarity. This region is rich in activity cliffs, and hence, provides the maximum amount of SAR information. However, compounds in this region tend to be refractory to the determination of reliable QSARs because functions describing QSARs of such compounds must be highly flexible to account for the high variability of their activity landscapes. Thus, a considerable amount of SAR data is required to ensure that the functions, which are highly non-linear, are adequately approximated. The upper left quadrant (III) corresponds to pairs of compounds with high activity similarity but low structure similarity. Compounds in this region exhibit relatively low *local* SAR information, although they do provide SAR information on *distributed* classes of active compounds (*Cf.* "scaffold hopping" [124, 125]). Because of the low similarities of compounds within this quadrant (*N.B.* that compounds in this set are typically dispersed in chemical space) it is generally not possible to develop reliable QSARs. However, if enough SAR data is available for compounds in both "active classes," it may be possible to develop QSARs for each class separately. The separate QSARs can then be merged into a single "distributed" QSAR. Care must be exercised in interpreting the points in quadrants (II) and (III), since high activity similarity obtains when pairs compounds have similar activities that can be high, moderate, low, or inactive. This can be indicated in SAS maps by distinguishing pairs of compounds using a color-coding scheme such as that described above for Fig. 12. Lastly, the lower left quadrant (IV) corresponds to pairs of compounds with both low activity and low structure similarity. This region contains very little if any SAR information and, thus, compounds in this region do not submit to QSAR or provide useful information on the nature of the activity landscape.

Figure 13 depicts the same SAS map shown in Fig. 12. Two points, one located in quadrant (II) and one in quadrant (III) are explicitly indicated by the green arrows that point from pairs of molecules located in Boxes (A) and (B), respectively. The point in quadrant (II) corresponds to an activity cliff since the activities of

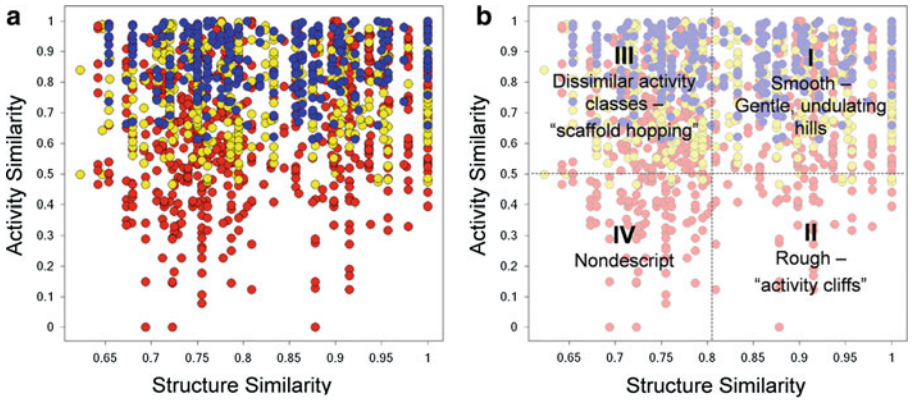


Fig. 12. Example of a structure–activity similarity (SAS) map (*Original figure provided courtesy of Dr. José Medina-Franco*). (**a**) Depiction of a SAS map for a prototypical dataset. Each data point indicates a pairwise comparison from a dataset of 51 compounds. Data points are color coded by the activity of the most active compound of a given pair. Red circles denote pairs where at least one compound is active; yellow circles indicate pairs where at least one compound is moderately active; blue circles indicates pairs where both compounds are inactive (or have low activity). Note that the scale for 'Structure Similarity' does not run from zero to unity. This is because of the limited size and diversity of the dataset. (**b**) Depiction of the four approximate quadrants of the SAS map (*see* text for further discussion).

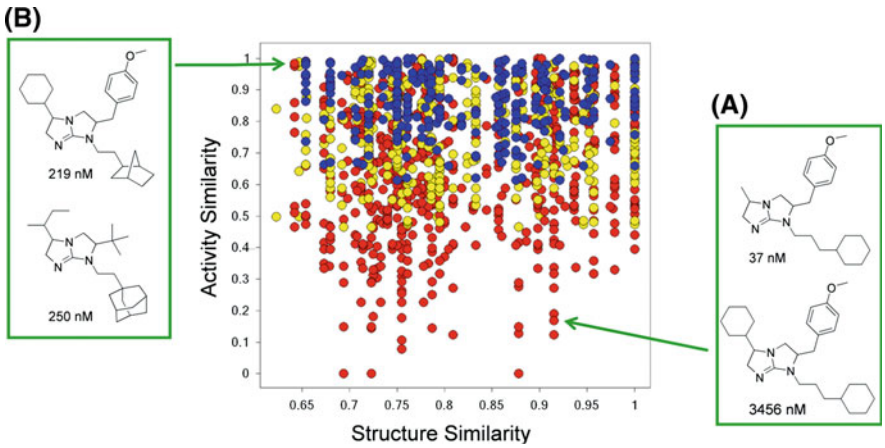


Fig. 13. SAS map given in Fig. 12 showing the structures of a pair of compounds involved in an activity cliff and a pair of compounds involved in scaffold hopping (*Original figure provided courtesy of Dr. José Medina-Franco*). The pair of compounds in Box (A) lie in quadrant (II), and the pair of compounds in Box (B) lie in quadrant (III) of the SAS map (*see* also Fig. 12b).

the two compounds in Box (A) differ by a factor of nearly 100, while their structure similarity is greater than 0.9. The two compounds associated with the designated point in quadrant (III) have moderate and nearly equal activities. However, since the dataset is of rather limited diversity – the smallest similarity between any two compounds in the set is 0.58 (*vide supra*) – it is not expected that the compounds associated with this datapoint will exhibit significant scaffold hopping, which is confirmed by the

structures in Box (B). It is expected that noteworthy scaffold hopping will require similarity values most likely of 0.40 or less. Thus, one might only expect this to occur in campaigns where relatively diverse sets of compounds are screened.

3.3.2. An Information-Theoretic Analysis of SAS Maps

Information theory provides a suitable framework for analyzing the information content of pairs of compounds located in the different quadrants of SAS maps, such as that depicted in Fig. 12b. Information (in "bits"), sometimes called "surprisal," is related to Shannon entropy [126] and is given by

$$\mathcal{I}(\mathrm{A}) = \log_2 \frac{1}{P(\mathrm{A})} = -\log_2 P(\mathrm{A}), \tag{82}$$

where $P(\mathrm{A})$is the probability of observing a specific activity (e.g., high, moderate, low, etc.) activity for molecule A. The condensed notation employed here is used for simplicity. Technically, $P(\mathrm{A})$ should *conditioned* on the activity of molecule B and the similarity, $S(\mathrm{A,B})$, of the two molecules. For example, what is the probability that molecule A is active (inactive) given that molecule B is active (inactive) and the similarity between them is high (low). Equation (82) makes sense from the following point of view, namely, the more likely an event is to be observed, the less information will be obtained in observing it, that is there is less "surprise" in observing it. For example, if an urn is filled with 90 red balls and 10 green balls, there is a 90% chance of drawing a red one and a 10% chance of drawing a green one. Thus, drawing a red ball is less surprising than drawing a green one and hence carries less information.

Although the number of exceptions to the rule that "similar compounds tend to possess similar activities" [4] is growing rapidly, it is not unreasonable to assume, at least as a working hypothesis, that this rule holds approximately. A compound located in the neighborhood of, say, an active compound is likely to also be active. If this is not the case, and the compound is found to be inactive, the result is surprising. Hence, such an observation carries higher information than if the compound was active as expected. Using this logic and the basics of information theory it is possible to *qualitatively* assess the degree of information possessed by pairs of compounds located within the different quadrants of a SAS map.

For example, consider two highly similar compounds chosen at random from a given chemical space. If one of the compounds is known to be active, then it is reasonable to expect that the other would have a high probability of also being active. This situation is exemplified by compounds residing in quadrant (I). Thus, although compounds in this quadrant are appropriate for the development of reliable QSARs, they provide little information, in the information-theoretic sense. If, on the other hand, one of the compounds is active but the other is inactive, a condition

exemplified by the activity cliffs associated with compounds in quadrant (II), this corresponds to a high information local case. However, due to the rapidly fluxuating nature of the activity landscape for compounds located in quadrant (II) it is difficult to construct meaningful QSARs (*see* Subsection 3.3.1 for further discussion). If two dissimilar compounds are chosen at random, and one is known to be active, it follows from the above discussion that the other is likely to be inactive. However, if it is shown to be active, a situation that obtains for compounds located in quadrant (III), this corresponds to a high-information case since a new class of compounds has been identified. Lastly, pairs of compounds that have low activity and structure similarity and are found in quadrant (IV) have little relationship to each other and, hence, have low information.

The above discussion can also be couched in terms of inactive compounds. This is obtained by interchanging the words "active" and "inactive." In such a case, inactive compounds in quadrants (I) and (III) have low information and are also not appropriate for QSAR studies. Table 1 provides a summary of the features of SAS maps.

3.3.3. Alternative Representations of Activity Landscapes

Many variants of SAS maps are possible. One that has received some attention recently is the multi-fusion similarity (MFS) maps developed by Medina-Franco et al. [127]. This work is a two-dimensional extension of the work carried out by Willet's group on similarity-based data fusion methods [25–28].

Several other approaches aimed at describing activity landscapes have been published [128, 129]. These approaches are

Table 1
Structure–activity similarity (SAS) maps

Quad	Similarity		Landscape	Cpd activity[a]	QSAR[b]	Information content
	Activity	Structure				
I	High	High	Gentle hills	High/high Low/low	+ –	Low
II	Low	High	Activity cliffs	High/low	–	High
III	High	Low	Multiple, separated active regions	High/high Low/low	–/+ –	High to moderate
IV	Low	Low	Nondescript	High/low	–	Very low

[a]If a pair of compounds has high activity similarity, both of the compounds could have high or low activities. This has been explicitly designated (*viz.*, High/High or Low/Low) for compounds in quadrants (I) and (III)

[b]The "+" sign indicates that it is possible to construct a reasonable QSAR; the "–" sign indicates that a QSAR is either not possible or problematic at best; the "–/+" indicates that it is a QSAR is not possible unless a significant amount of data is available for both of the dissimilar activity classes

based on two indices – SALI [128] and SARI [129]. The former, which is defined as

$$\mathrm{SALI}(\mathrm{A},\mathrm{B}) = \frac{|\mathrm{Act}(\mathrm{A}) - \mathrm{Act}(\mathrm{B})|}{1 - S(\mathrm{A},\mathrm{B})}, \tag{83}$$

is designed to identify the presence of activity cliffs between pairs of compounds. Thus, it provides a *local* characterization of the neighborhoods surrounding activity cliffs. A more global view is obtained by stitching together compounds using a directed graph-theoretical formalism. In this formalism, the nodes represent compounds. A *directed* edge is drawn between two compounds and points towards the higher activity one depending upon whether the SALI index is above a given threshold value. This representation provides a graphical means for portraying neighborhood as well as global relationships among compounds associated with activity cliffs – as the SALI threshold value is lowered a more complete picture of the inter-relationships among activity cliff regions emerges.

In contrast, the SARI index takes a more global approach to activity landscapes. It is based on the average of the "continuity" score, $\langle \mathrm{Score}_{\mathrm{cont}} \rangle$, and the "discontinuity" score $\langle \mathrm{Score}_{\mathrm{discont}} \rangle$, The former is computed by taking the potency-weighted sum of all pairwise dissimilarities and, thus, is a measure of the potency and diversity of the set of compounds under consideration. In contrast, the latter is computed by taking similarity-weighted average potency among pairs of compounds that exceed a given similarity threshold value. Large values of the discontinuity score indicate the presence of activity cliffs. After normalizing both scores to the unit interval, they are combined as shown in Eq. (84) to yield the SARI index value,

$$\mathrm{SARI} = \tfrac{1}{2}\left[\langle \mathrm{Score}_{\mathrm{cont}} \rangle_{\mathrm{norm}} + \left(1 - \langle \mathrm{Score}_{\mathrm{discont}} \rangle_{\mathrm{norm}}\right)\right]. \tag{84}$$

High SARI values correspond to predominantly continuous activity landscapes, while low values correspond to predominantly discontinuous landscapes. Intermediate values correspond to mixed landscapes. Hence, the SARI index provides a more global measure of the activity landscape than does the SALI index. However, it provides a much less detailed picture of the local environments in an activity landscape, although a local discontinuity index has also been defined for the former [129]. Subsequent work in Bajorath's laboratory [130] has extended their SARI-based approach using network-like similarity graphs (NSG) along with local information to provide a more detailed picture of activity landscapes.

Lastly, since similarity values are influenced by the representation used to encode the molecular information, it is desirable to develop a method that is less sensitive to representation. To

address this difficult and persistent problem, Medina-Franco *et al.* [131] developed the concept of *consensus activity cliffs*. Consensus activity cliffs, which are obtained by analyzing multiple activity landscapes generated by different similarity methods, are characterized by the Degree of Consensus (DoC) between any pair of similarity methods. Using this approach, the authors were able to identify activity cliffs of improved reliability that persisted with respect to a number of similarity methods.

3.4. A General Similarity-Based Approach for Representing Chemical Spaces

Vector-based representations of chemical spaces are quite common in cheminformatics. In the usual molecular fragment-based approach the coefficients of the vector components are generally binary- or positive integer-valued (*see* Subsections 2.2.1 and 2.2.3). In continuous vector representations (*see* Subsection 2.3), on the other hand, the vector coefficients are typically associated with the values of atomic and molecular properties. All of these representations can be used to describe chemical spaces, either directly or in terms of their related molecular similarities, proximities, or distances.

In the following, a general similarity-based approach for constructing continuous vector representations of molecules is presented that is based on what might be called "molecular basis vectors" instead of molecular fragments or properties (*Cf.* Subsection 2.3). The method is reminiscent of those in molecular quantum mechanics [132] that employ atomic orbitals or group functions as a basis for describing whole molecules. A detailed account with a number of examples can be found in a recent work by Raghavendra and Maggiora [95]. As will be seen in the sequel, the key to the method is an *ansatz* that equates the inner product between a pair of molecular vectors to their corresponding similarities. The method's power resides in the fact that any reasonable similarity can be used and that the detailed nature of the molecular vector need not be known since it is not required for the computation of similarity. This situation is reminiscent of that in many kernel learning methods [133–135] where elements of the Gramian matrix, which are analogous to the similarities used here, can be determined indirectly without knowledge of the explicit form of the vectors in the underlying feature space.

Choose a set of p molecules as a molecular basis for representing the chemical space of interest

$$\mathrm{B} = \{b_1, b_2, ..., b_p\}, \tag{85}$$

which can be written as the "row vector" (i.e. a $1 \times n$-dimensional matrix)

$$\mathrm{B} = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_p), \tag{86}$$

where the $\mathbf{b}_i$, $i = 1, 2, ..., p$ correspond to *molecular basis vectors*. Here, the word "vector" refers to an abstract object with direction

and magnitude located at the origin of an appropriate coordinate system and satisfying the multiplicative and additive properties of a linear vector space [36]. Such objects, which are depicted in a lower-case, bold-face Arial font, e.g. "$\mathbf{b}_i$, $\mathbf{m}_k$, ...", are *basis-set independent*. Component vectors, which are *basis-set dependent*, are depicted in a lower-case, bold-face Times New Roman font, e.g. "$\mathbf{v}_i$, $\mathbf{m}_k$, ...". Their components are depicted in lower-case italic type and are grouped together in $n \times 1$ column matrices (*see* Eq. (91)).

Consider the similarities of all of the elements of the molecular basis set with respect to each other. This formally generates the matrix of inner products [136]

$$\mathbf{S} = \langle \mathbf{B}, \mathbf{B} \rangle = \begin{pmatrix} \langle \mathbf{b}_1, \mathbf{b}_1 \rangle & \langle \mathbf{b}_1, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_1, \mathbf{b}_p \rangle \\ \langle \mathbf{b}_2, \mathbf{b}_1 \rangle & \langle \mathbf{b}_2, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_2, \mathbf{b}_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{b}_p, \mathbf{b}_1 \rangle & \langle \mathbf{b}_p, \mathbf{b}_2 \rangle & \cdots & \langle \mathbf{b}_p, \mathbf{b}_p \rangle \end{pmatrix}, \tag{87}$$

$$\mathbf{S} = \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p,1} & S_{p,1} & \cdots & S_{p,p} \end{pmatrix}. \tag{88}$$

Since the off-diagonal elements are, in general, non-zero and since the diagonal elements are of unit value the molecular basis vectors constitute a set of non-orthogonal unit vectors. A crucial feature of this approach is that the exact nature of the molecular basis vectors need not be known. Due to the *ansatz*, only their inner products are required, and these are taken to be equivalent to the similarities among pairs of the corresponding molecules (*vide supra*). Since $0 < S_{i,j} \leq 1$, the elements of **S** are analogous to the basis-set overlaps integrals familiar in quantum chemistry [69]. Moreover, **S** is positive definite if all of the elements of the molecular basis are linearly independent; when they are not, **S** becomes positive semi-definite. Hence, the definiteness of the **S** matrix provides a measure of the linear independence of the molecular basis set.

Similarities between molecular basis elements can be evaluated in a number of different ways. For example, suppose the $\mathbf{b}_i \Leftrightarrow \mathrm{G}_i$, that is the i-th element of the molecular basis is a *labeled chemical graph* of a molecule. Then $S_{i,j} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle \equiv S_{\mathrm{Tan}}(G_i, G_j)$, where the Tanimoto similarity, S_{Tan}, is evaluated as in Eq. (15). The set of labeled graphs $G = \{G_1, G_2, ..., G_p\}$ can be referred to as a "chemical graph basis". Similarities can also be computed using a bit-vector representation or from their 3-D structures or molecular fields as described in the preceding sections. In some

applications **S** is equivalent to what is typically called the *metric matrix*; in statistics **S** is equivalent to the *correlation matrix* [136].

Consider a given molecule m_i within a set of n molecules

$$\mathrm{M} = \{m_1, m_2, ..., m_n\}. \tag{89}$$

A molecule $m_i \in \mathrm{M}$ generally does not correspond to any of the basis molecules in B, although such a correspondence is not specifically precluded on mathematical grounds because of the non-orthogonality of the molecular basis. In matrix notation, a given molecule $m_i \in \mathrm{M}$, can be represented as a abstract vector m_i in the molecular basis,

$$\mathbf{m}_i = \mathbf{B}\,\mathbf{m}_i, \tag{90}$$

where the column vector of coefficients is given by

$$\mathbf{m}_i = \begin{pmatrix} m_i(b_1) \\ m_i(b_2) \\ \vdots \\ m_i(b_p) \end{pmatrix}. \tag{91}$$

To compute the various cosine-like similarity indices it is necessary to evaluate the inner product $\langle \mathbf{m}_i, \mathbf{m}_j \rangle$ and vector norm $\| \mathbf{m}_i \| = \sqrt{\langle \mathbf{m}_i, \mathbf{m}_i \rangle}$ (*see* also Eqs. (45) and (46)):

$$\begin{aligned} \langle \mathbf{m}_i, \mathbf{m}_j \rangle &= \langle \mathbf{B}\,\mathbf{m}_i,\ \mathbf{B}\,\mathbf{m}_j \rangle \\ &= \mathbf{m}_\mathbf{i}^{\mathrm{T}} \langle \mathbf{B},\ \mathbf{B} \rangle\, \mathbf{m}_j \end{aligned}. \tag{92}$$

In expanded form, the inner product is given by

$$\begin{aligned} \langle \mathbf{m}_i, \mathbf{m}_j \rangle = &\left(m_i(b_1), m_i(b_2), ..., m_i(b_p) \right) \\ &\times \begin{pmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,p} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p,1} & S_{p,1} & \cdots & S_{p,p} \end{pmatrix} \begin{pmatrix} m_i(b_1) \\ m_i(b_2) \\ \vdots \\ m_i(b_p) \end{pmatrix}. \end{aligned} \tag{93}$$

In summation form, Eq. (93) becomes

$$\langle \mathbf{m}_i, \mathbf{m}_j \rangle = \sum_{k=1}^{p} \sum_{\ell=1}^{p} m_i(b_k) \cdot m_j(b_\ell) \cdot S_{k,\ell}, \tag{94}$$

where $S_{i,i} = 1$ for $i = 1, 2, ..., p$. Comparing Eq. (94) with Eq. (45) shows that the elements of the S-matrix modulate the product of the vector components and the cross-terms, "$m_i(b_k) \cdot m_j(b_\ell)$", remain. When the basis is orthonormal $\mathbf{S} = \mathbf{I}$ and Eq. (94) reduces to Eq. (45). Similarly, the vector norm for the i-th molecule is given by

$$\|\mathbf{m}_i\| = \sum_{k=1}^{p} \sum_{\ell=1}^{p} m_i(b_k) \cdot m_i(b_\ell) \cdot S_{k,\ell}, \tag{95}$$

which reduces to Eq. (46) when $\mathbf{S} = \mathbf{I}$. These relationships clearly show the important role played by the metric matrix **S**. Since the various cosine-like similarity indices all depend on the quantities given in Eqs. (94) and (95), it follows that these indices also depend upon **S**, but this dependence is routinely neglected in most calculations. Euclidean (*see* Eq. (42)) and other distances are likewise affected by the metric matrix:

$$\begin{aligned} d_{\text{Euc}}(\mathbf{m}_i, \mathbf{m}_j) &= \left\| \mathbf{m}_i - \mathbf{m}_j \right\| \\ &= \sqrt{\left\langle (\mathbf{m}_i - \mathbf{m}_j), (\mathbf{m}_i - \mathbf{m}_j) \right\rangle} \\ &= \sqrt{\sum_{k=1}^{p} \sum_{\ell=1}^{p} \left(m_i(b_k) - m_j(b_\ell) \right) \cdot \left(m_i(b_k) - m_j(b_\ell) \right) \cdot S_{k,\ell}} \end{aligned} . \tag{96}$$

As was true in the two cases above, Eq. (96) reduces to Eq. (42) in an orthonormal basis.

There are numerous ways in which to orthonormalize a basis [137]. Here, we choose to employ the *symmetric orthonormalization* procedure described by Löwdin [136]

$$\bar{\mathbf{B}} = \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}}, \tag{97}$$

where

$$\bar{\mathbf{B}} = (\bar{\mathbf{b}}_1, \bar{\mathbf{b}}_2, \ldots, \bar{\mathbf{b}}_p). \tag{98}$$

This has the benefit over other orthogonality procedures that the new basis is as close as possible, in a least square sense, to the original basis [138]. Computing the inner product, $\left\langle \bar{\mathbf{B}}, \bar{\mathbf{B}} \right\rangle = \left\langle \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}}, \mathbf{B}\,\mathbf{S}^{-\frac{1}{2}} \right\rangle = \mathbf{S}^{-\frac{1}{2}} \langle \mathbf{B}, \mathbf{B} \rangle\, \mathbf{S}^{-\frac{1}{2}} = \mathbf{S}^{-\frac{1}{2}}\,\mathbf{S}\,\mathbf{S}^{-\frac{1}{2}} = \mathbf{I}$, shows that the basis is indeed orthonormal.

Right multiplying the terms in Eq. (97) by $\mathbf{S}^{\frac{1}{2}}$ gives $\mathbf{B} = \bar{\mathbf{B}}\,\mathbf{S}^{\frac{1}{2}}$, which upon substitution into Eq. (90) yields

$$\begin{aligned} \mathbf{m_i} &= \mathbf{B}\,\mathbf{m}_i \\ &= (\bar{\mathbf{B}}\,\mathbf{S}^{\frac{1}{2}})\mathbf{m}_i = \bar{\mathbf{B}}\,(\mathbf{S}^{\frac{1}{2}}\,\mathbf{m}_i) \\ &= \bar{\mathbf{B}}\,\bar{\mathbf{m}}_i \end{aligned} , \tag{99}$$

where the "expansion coefficients" (i.e. components) in the new, orthonormal basis are given by

$$\bar{\mathbf{m}}_i = \mathbf{S}^{\frac{1}{2}}\,\mathbf{m}_i. \tag{100}$$

As was the case for the basis above, Eq. (100) can be rearranged to give the expansion coefficients in the original, non-orthogonal basis,

$$\mathbf{m}_i = \mathbf{S}^{-\frac{1}{2}}\,\mathbf{m}_i. \tag{101}$$

Thus, this equation provides the means for determining the components of $\mathbf{m}_i$ in the original basis given the components in the orthonormal basis, which are easily determined. This can be accomplished by first taking the inner product of the k-th orthonormal basis element with $\mathbf{m}_i$ (*see*, e.g., Eq. (94))

$$\begin{aligned}\left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle &= \left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{B}}\, \bar{\mathbf{m}}_i \right\rangle \\ &= \sum_{\ell=1}^{p} \left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{b}}_\ell \right\rangle \bar{m}_i(\bar{b}_\ell).\end{aligned} \tag{102}$$

Since $\left\langle \bar{\mathbf{b}}_k, \bar{\mathbf{b}}_\ell \right\rangle = \delta_{k,\ell}$, where the Kronecker delta, $\delta_{k,k} = 1$ and $\delta_{k,\ell} = 0$ for $k \neq \ell$ the k-th component of $\mathbf{m}_i$ is given by

$$\bar{m}_i(\bar{b}_k) = \left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle \quad \text{for } k = 1, 2, ..., p. \tag{103}$$

Because $\mathbf{m}_i$ is normalized with respect to the Euclidean norm

$$\sum_{k=1}^{p} \bar{m}_i(\bar{b}_k)^2 = 1 \tag{104}$$

the *square* of each component value, $\bar{m}_i(\bar{b}_k)$, gives the fraction of the molecule represented by its corresponding orthonormal basis element $\bar{\mathbf{b}}_k$.

To evaluate the inner product $\left\langle \bar{\mathbf{b}}_k, \mathbf{m}_i \right\rangle$, $\bar{\mathbf{b}}_k$ must be expanded in terms of the original non-orthogonal basis (*see* Eq. (97)), that is

$$\bar{\mathbf{b}}_k = \sum_{\ell=1}^{p} \mathbf{b}_\ell\, S_{k,\ell}^{-\frac{1}{2}}. \tag{105}$$

Substituting Eq. (105) into Eq. (103) yields

$$\bar{m}_i(\bar{b}_k) = \sum_{\ell=1}^{f} S_{k,\ell}^{-\frac{1}{2}} \left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle. \tag{106}$$

The inner-product terms $\left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle$ can now be evaluated in exactly the same manner as was described earlier using the chosen similarity measure. For example, $\left\langle \mathbf{b}_\ell, \mathbf{m}_i \right\rangle = \mathrm{S_{Tan}}(\mathrm{G}_\ell, \mathrm{G}_m)$, where "$\mathbf{b}_\ell$" is the labeled graph corresponding to ℓ-th basis molecule, "$\mathbf{m}_i$" is the labeled graph corresponding to the i-th molecule, and $S_{\mathrm{Tan}}(G_\ell, G_m)$ is the chemical graph-theoretical Tanimoto similarity coefficient.

This approach can, in many instances, be extended even to cases where the basis is comprised of physico-chemical, topological, or other such parameters. The similarity matrix is replaced in these cases by the correlation matrix computed with respect to the "basis set" of parameters.

Agrafiotis et al. [139], developed a similar approach to generate vectors for input into neural nets. Although these authors did not account for the inherent non-orthogonality of the "basis," in their work, the issue of the orthogonality of the basis may be less critical than it is here, and the mappings they generated seem to

be sufficiently stable. Another related approach comes from Villar and co-workers [140]. In this case, however, the basis consisted of a set of proteins. The interaction of each molecule in the training set to each of the proteins in the "basis proteins" was measured experimentally, and the expansion coefficients were determined using a least squares procedure. Again, non-orthogonality of the basis was not explicitly addressed, although the choice of the basis proteins did involve an assessment of correlations among them.

Randic [141–143] has investigated the role of orthogonalized descriptors in multivariate regressions. In his work he points out that although the predictions obtained with orthogonal or non-orthogonal descriptors are the same, the stability of the regression coefficients is much greater in the former case. Also, adding a new, orthogonal descriptor to set of orthogonal descriptors does not affect the values of the previously determined regression coefficients. This is definitely not the case for non-orthogonal descriptors where addition of a new descriptor can cause all of the coefficients to fluctuate significantly depending on the degree of collinearity of the new descriptor with those in the original set.

4. Summary and Conclusions

This chapter provides an overview of the mathematics that underlies many of the similarity measures used in cheminformatics. Each similarity measure is made up of two key elements: (1) A mathematical representation of the relevant molecular information and (2) some form of similarity measure, index or coefficient that is compatible with the representation. The mathematical forms typically used are sets, graphs, vectors, and functions, and each is discussed at length in this chapter.

As was described in Subsection 2.1, chemical graphs are a subclass of mathematical graphs, and thus many of the features of the latter can be taken over to the former. A number of graph metrics, such as the size of a graph and the distance between two graphs, have been applied to chemical graphs. In addition, similarity measures, such as the Tanimoto similarity index, also have their corresponding graph-theoretical analogs and have been used in a number of cases, albeit on relatively small sets of molecules. Although chemical graphs are the most familiar and intuitive representation of molecular information to chemists, they have been used relatively rarely in MSA. This is due primarily to computational difficulties brought on by the need to evaluate the MCS, an NP-complete computational problem that is required by most graph-based distance and similarity measures.

Subsection 2.2 describes the properties of discrete-valued feature vectors, with components given by finite, ordered sets of

values. The most prevalent class is that of vectors with binary-valued components, which are mathematically equivalent to classical sets. Here features are either in the set (component value of "1") or not in the set (component value of "0"). Because we are essentially dealing with sets, the distance and similarity measures used are typically related to set measures (i.e. cardinalities) and not to the types of inner (scalar) products defined on linear vector spaces. A hypercubic mathematical space associated can be associated with classical sets, where the dimension of the space is equal to the number of elements in the universal set and each vertex of the hypercube corresponds to a subset, including the null and universal sets. Distances in these spaces are appropriately Hamming distances that satisfy an ℓ_1 metric. Although Euclidean distances are sometimes used, they are inappropriate in such hypercubic spaces. Most similarity indices are taken to be symmetric ("A is as similar to B as B is similar to A"), but Tversky defined an infinite family of asymmetric indices related to the Tanimoto similarity index, some of which may be useful for similarity-related tasks such as similarity searching.

Another class of discrete-valued feature vectors useful in MSA is integer- and categorical-valued feature vectors. Here, the vectors are mathematically equivalent to multisets and not directly to classical sets, although multisets can be reformulated as classical sets. The components of the vectors now indicate the number of times a given feature occurs or the ordered set of categorical values corresponding to the given feature or property. Although care must be taken, distance and similarity measures analogous to those used for binary-valued vector components can be used here as well.

In Subsection 2.3, the important class of vectors with continuous-valued components is described. A number of issues arise in this case. Importantly, since the objects of concern here are vectors, the mathematical operations employed are those applied to vectors such as addition, multiplication by a scalar, and formation of inner products. Care, however, must be exercised because in some cases the "vector objects" may not reside in linear vector spaces. While distances between vectors are used in similarity studies, inner products are the most prevalent type of terms found in MSA. Such similarities, usually associated with the names Carbó and Hodgkin, are computed as ratios, where the inner product term in the numerator is normalized by a term in the denominator that is some form of mean (e.g. geometric or arithmetic) of the norms of the two vectors.

The notion of an orthogonal set of "basis vectors" is also of significance here and is particularly important since as discussed in Subsection 2.3.2 it is in many instances ignored. In a non-orthogonal basis the associated similarity matrix defines the metric of the space in which the vectors "live." Thus, "measurements" such as

the distance or the angle between two vectors in the space are dependent on the metric of that space. Further discussion on this point is presented in Subsection 3.1.3 that describes a general approach for dealing with non-orthogonal bases and explores some of the consequences of ignoring non-orthogonality in the description of chemical spaces. While most of the discussion deals with what are called "molecular basis sets", the method can also deal with physico-chemical, topological, or other such descriptors. However, in these cases the correlation matrix replaces the similarity matrix.

Subsection 2.4 addresses the use of field-based functions in MSA. Field-based functions, which can be thought of as infinite-dimensional vectors, are used primarily in 3-D MSA. Here, molecular fields (e.g. steric or electrostatic) or pseudo-fields (e.g. lipophilic) of the molecules being compared are matched, using various similarity measures, the most popular being those of Carbó or Hodgkin. Because 3-D field-based similarities are non-linear functions, multiple solutions corresponding to different alignments are possible. This has raised the issue of how one obtains consistent multi-molecule consistent alignments, a subject that is treated in Subsection 2.4.3. Conformational flexibility adds a new degree of difficulty to studies of 3-D MSA, and this has been dealt with in a number of ways. The most widespread approach is by standard conformational analysis. Since such an analysis leads to many conformations clustering is usually used to group the conformations as a basis for identifying a smaller set of prototypical conformations. Molecular similarity is then carried out by pairwise matching the fields generated by each conformational prototype in one molecule with each conformational prototype in the other molecule being compared. This represents a rather substantial computational problem that has been ameliorated somewhat using Fourier transforms to separate translational from rotational motions in the optimization process. Alternatively, several procedures have been developed that combine conformational analysis with 3-D similarity matching simultaneously in the optimization process. Both approaches are, however, computationally demanding, although the latter is somewhat better in this regard. Since multiple conformers for each molecule may contribute to the overall similarity, Subsection 2.4.5 deals with a possible way of combining this information into a single multi-conformer dependent similarity.

Subsection 2.5 provides a very brief discussion of molecular dissimilarity measures that are basically the complement of their corresponding molecular similarity measures. This section also presents reasons as to why similarity is preferred over dissimilarity, except in studies of diversity, as a measure of molecular resemblance.

The concept of chemical space pervades, either explicitly or implicitly, much of the literature in cheminformatics. As is

discussed in Subsection 3, chemical spaces are induced by various similarity measures. The different similarity measures do not necessarily give rise to topologically equivalent chemical spaces – nearest-neighbor relations are generally not preserved among chemical spaces induced by different similarity measures. The consequences of this are manifold. An especially egregious consequence is that the results of similarity searches based upon different similarity measures tend can differ substantially. And there is no easy solution to this problem.

Chemical spaces fall into two broad categories, coordinate-based and coordinate-free. Coordinate-based chemical spaces, even those of relatively low dimensionality, tend to be of difficult to visualize directly. Coordinate-free chemical spaces cannot be visualized directly since coordinates do not exist. In both cases it is possible to develop reduced-dimension representations that are easier to work with theoretically and also afford possibilities for visualization. Constructing reduced dimension spaces, which is discussed in Subsection 3.2 for the case of chemical spaces, is a difficult problem that pervades many fields, and methods developed in these fields have proved useful in cheminformatics, albeit to varying degrees.

The growing importance of activity landscapes, and more specifically activity cliffs, are the subject of Subsection 3.3. A growing body of data shows quite clearly that the old aphorism "similar molecules have similar activities" is no longer entirely valid due to the presence of activity cliffs. This has important implications for QSAR studies. However, considerable SAR information is contained in activity cliffs that arise when small changes in similarity lead to large changes in activity. Recently, their growing importance is being recognized, and a number of studies addressing many issues associated with them have been published.

MSA has developed substantially over the years, especially as digital computers became faster, more compact, and widely available to scientists. Handling large sets of molecules is generally not a problem. The main problem confronting MSA is the problem of the lack of topological invariance of the chemical spaces induced by the various similarity measures. Unfortunately, this problem may be fundamentally related to the inherent subjectivity of similarity and thus cannot be addressed in any simple manner.

5. Appendix: A New Notation for Classical Sets

Sets are very general mathematical objects that are used in many branches of mathematics. Here the focus is on *finite* sets, that is sets with a finite set of elements. A key concept in set theory is that of the *universal set*, U, sometimes called the *universe of*

discourse, which is an unordered collection of n elements $x_1, x_2, ..., x_k, ..., x_n$ and is given by

$$\mathrm{U} = \{x_1, x_2, ..., x_k, ..., x_n\}. \tag{107}$$

All sets in this "universe," including U and the null or empty set Ø, are subsets of U. A subset A is typically written as, for example,

$$\mathrm{A} = \{x_1, x_3, x_9, x_{12}, x_{17}, x_{18}, ...\}, \tag{108}$$

but his notation can become awkward and cumbersome for large, complex sets. A more general and powerful notation, which utilizes the concept of an *indicator* or *characteristic function*, $A(x_k)$, is illustrated in Eq. (109),

$$\mathrm{A} = \{A(x_1), A(x_2),..., A(x_k),..., A(x_n)\}, \tag{109}$$

where $A(x_k)$ characterizes the membership of each element in the set is given by

$$A(x_k) = \begin{cases} 1 & \text{if } x_k \in A \\ 0 & \text{if } x_k \notin A \end{cases}. \tag{110}$$

Thus, in the universal set $A(x_k) = 1 \quad \text{for } k = 1, 2, ..., n$, that is all elements of the universal set have a membership-function value of unity.

Note that this representation differs from that usually used (*see* Eq. (108)) where only those elements actually in the set, that is those elements for which $A(x_k) = 1$, are included explicitly. All possible sets A, including the empty and universal sets Ø and U, are subsets of U, i.e. $A \subseteq U$. While this notation may be unfamiliar, it is completely equivalent to that used for binary vectors or "bit vectors." Fuzzy sets, although not treated in this chapter, can also be represented in this notation with the modification that elements of the set are no longer confined to the binary values {0,1}; fuzzy sets can take on all values between and including zero and unity [51]. A number of useful operations between two sets, A and B, are given in the notation introduced above:

$$A \cap B = \min_k [A(x_k), B(x_k)] \quad \text{Set Intersection} \tag{111}$$

$$A \cup B = \max_k [A(x_k), B(x_k)] \quad \text{Set Union} \tag{112}$$

$$\begin{aligned} A^{\mathrm{c}} &= \{1 - A(x_1), 1 - A(x_2), ..., 1 - A(x_n)\} \\ &= \{A^{\mathrm{c}}(x_1), A^{\mathrm{c}}(x_2), ..., A^{\mathrm{c}}(x_n)\} \end{aligned} \quad \text{Set Complementation} \tag{113}$$

$$\begin{aligned} A - B = A \cap B^{\mathrm{c}} &= \min_k [A(x_k), 1 - B(x_k)] \\ &= \min_k [A(x_k), B^{\mathrm{c}}(x_k)] \end{aligned} \quad \text{Set Difference} \tag{114}$$

$$A \subseteq B = A(x_k) \leqslant B(x_k) \text{ for all } k \quad \text{Subsethood} \tag{115}$$

$$|\mathrm{A}| = \sum_k A(x_k) \quad \text{Cardinality} - \text{Set} \tag{116}$$

$$|A \cap B| = \sum_k \min[A(x_k), B(x_k)] \quad \text{Cardinality} - \text{Set Intersection} \tag{117}$$

$$|A \cup B| = \sum_k \max[A(x_k), B(x_k)] \quad \text{Cardinality} - \text{Set Union} \tag{118}$$

$$|A| = |A - B| + |A \cap B| \quad \text{Cardinality} - \text{Set} \tag{119}$$

$$|A \cup B| = |A| + |B| - |A \cap B| \quad \text{Cardinality} - \text{Set Union} \tag{120}$$

$$|A \cup B| = |A - B| + |B - A| + |A \cap B| \quad \text{Cardinality} - \text{Set Union} \tag{121}$$

Relations, which are also sets, play an important role in set theory and in the similarity theory, but due to space limitations are not formally considered in this work.

Acknowledgments

The authors would like to thank Tom Doman for his constructive comments on the original version of this manuscript, and Mark Johnson, Mic Lajiness, John Van Drie, and Tudor Oprea for helpful discussions. Special thanks are given to Jurgen Bajorath and Jose Medina-Franco, for providing several figures and for their helpful comments.

References

1. Rouvray, D. (1990) The evolution of the concept of molecular similarity. In *Concepts and Applications of Molecular Similarity*, M.A. Johnson and G.M. Maggiora, Eds., Wiley, New York, Chapter 2.
2. Sheridan, R.P. and Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discovery Today* **7**, 903–911.
3. Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems*. Research Studies Press, Letchworth.
4. Johnson, M.A. and Maggiora, G.M., Eds. (1990) *Concepts and Applications of Molecular Similarity*. Wiley, New York.
5. Dean, P.M., Ed. (1994) *Molecular Similarity in Drug Design*. Chapman & Hall, Glasgow.
6. Tversky, A. (1977) Features of similarity. *Pyschol. Rev.* **84**, 327–352.
7. Chen, X. and Brown, F.K. (2007) Asymmetry of chemical similarity. *Chem. Med. Chem.* **2**, 180–182.
8. Willett, P., Barnard, J.P., and Downs, G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996.
9. Bender, A. and Glen, R.C. (2004) Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2**, 3204–3218.
10. Johnson, M.A. (1989) A review and examination of mathematical spaces underlying molecular similarity analysis. *J. Math. Chem.* **3**, 117–145.
11. Borg, I. and Groenen, P. (1997) *Modern Multidimensional Scaling*. Springer, New York.
12. Jolliffe, I.T. (2002) *Principal Component Analysis (Second Edition)*. Springer, New York.

13. Domine, D., Devillers, J., Chastrette, M., and Karcher, W. (1993). Non-linear mapping for structure-activity and structure-property modeling. *J. Chemometrics* **7**, 227–242.
14. Rush, J.A. (1999) Cell-based methods for sampling high-dimensional spaces. In *Rational Drug Design*, Truhlar, D.G., Howe, W.J., *et al.*, Eds., Springer, New York, pp. 73–79.
15. Rohrbaugh, R.H. and Jurs, P.C. (1987) Descriptions of molecular shape applied in studies of structure/activity and structure/property relationships. *Anal. Chim. Acta* **199**, 99–109.
16. Verloop, A. (1987) *The STERIMOL Approach to Drug Design*. Marcel Dekker, New York.
17. Mulliken, R.S. (1955) Electronic population analysis on LCAO-MO molecular wave functions. I. *J. Chem. Phys.* **23**, 1833–1840.
18. Stanton, D.T.; Jurs, P.C. (1990) Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Anal. Chem.* **62**, 2323–2329.
19. Kier, L.B. (1989) An index of molecular flexibility from kappa shape attributes. *Quant. Struct.-Act. Relat.* **8**, 221–224.
20. Kvasnička, V. and Pospíchal, J. (1989) Two metrics for a graph-theoretical model of organic chemistry. *J. Math. Chem.* **3**, 161–191.
21. Kvasnička, V. and Pospíchal, J. (1991) Chemical and reaction metrics for graph-theoretical model of organic chemistry. *J. Mol. Struct. (Theochem.)* **227**, 17–42.
22. Randić, M. (1992) Representation of molecular graphs by basic graphs. *J. Chem. Inf. Comput. Sci.* **32**, 57–69.
23. Baskin, I.I., Skvortsova, M.I., Stankevich, I.V., and Zefirov, N.S. (1995) On the basis of invariants of labeled molecular graphs. *J. Chem. Inf. Comput. Sci.* **35**, 527–531.
24. Skvortsova, M.I., Baskin, I.I., Stankevich, I.V., Palyulin, V.A., and Zefirov, N.S. (1998) Molecular similarity. I. Analytical description of the set of graph similarity measures. *J. Chem. Inf. Comput. Sci.* **38**, 785–790.
25. Ginn, C.M.R., Willett, P., and Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Perspec. Drug Disc. Design* **20**, 1–16.
26. Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., and Schuffenhauer, A. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44**, 1177–1185.
27. Whittle, M., Gillet, V.J., Willett, P., Alexander, A., and Loesel, J. (2004) Enhancing the effectiveness of virtual screening by fusing nearest-neighbor lists: A comparison of similarity coefficients. *J. Chem. Inf. Comput. Sci.* **44**, 1840–1848.
28. Whittle, M., Gillet, V.J., Willett, P., and Loesel, J. (2006) Analysis of data fusion methods in virtual screening: Similarity and group fusion. *J. Chem. Inf. Model.* **46**, 2206–2219.
29. Mestres, J., Rohrer, D.C., and Maggiora, G.M. (1999) A molecular-field-based similarity study of non-nucleoside HIV-1 reverse transcriptase inhibitors. *J. Comput.-Aided Mol. Design* **13**, 79–93.
30. Trinajstić, N. (1992) *Chemical Graph Theory*. CRC Press, Boca Raton, Florida.
31. Harary, F. (1969) *Graph Theory*. Addison-Wesley Publishing Company, Reading, Massachusetts.
32. Raymond, J.W. and Willett, P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Design* **16**, 521–533.
33. Mason, J.S., Morize, I., Menard, P.R., Cheney, D.L., Hulme, C., and Labaudiniere, R.F. (1999) New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med Chem.* **42**, 3251–3264.
34. Devillers, J. and Balaban, A.T., Eds. (1999) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, Amsterdam, The Netherlands.
35. Pearlman, R.S. and Smith, K.M. (1998) Novel software tools for chemical diversity. *Perspec. Drug Disc. Design* **9/10/11**, 339–353.
36. Halmos, P.R. (1958) *Finite-Dimensional Vector Spaces, Second Edition*. D. Van Nostrand Company, Inc., Princeton, New Jersey.
37. Mestres, J., Rohrer, D.C., and Maggiora, G.M. (1997) MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **18**, 934–954.
38. Thorner, D.A., Willett, P., Wright, P.M., and Taylor, R. (1997) Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs. *J. Comput.-Aided Mol. Design* **11**, 163–174.

39. Du, Q., Arteca, G.A., and Mezey, P.G. (1997) Heuristic lipophilicity potential for computer-aided rational drug design. *J. Comput.-Aided Mol. Design* **11**, 503–515.
40. Oden, J.T. and Demkowicz, L.F. (1996) *Applied Functional Analysis.* CRC Press, Boca Raton, Florida.
41. Petke, J.D. (1993) Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem.* **14**, 928–933.
42. Cramer, R.D., Patterson, D.E., and Bunce, J.D. (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Amer. Chem. Soc.*, **110**, 5959–5967.
43. Bandemer, H. and Näther, W. (1992) *Fuzzy Data Analysis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.
44. Kaufmann, A. and Gupta, M.M. (1985) *An Introduction to Fuzzy Arithmetic – Theory and Applications.* Van Nostrand Reinhold, New York.
45. McGregor, J. and Willett, P. (1981) Use of a maximal common subgraph algorithm in the automatic identification of the ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.* **21**, 137–140.
46. Johnson, M. (1985) Relating metrics, lines, and variables defined on graphs to problems in medicinal chemistry. In *Graph Theory and its Applications to Algorithms and Computer Science*, Y. Alavi *et al.*, Eds., Wiley, New York, pp.457–470.
47. Hagadone, T.R. (1992) Molecular substructure similarity searching: Efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* **32**, 515–521.
48. Rusinko, A., Farmen, M.W., Lambert, C.G., and Young, S.S. (1997) SCAM: Statistical classification of activities of molecules using recursive partitioning. 213th ACS Natl. Meeting, San Francisco, CA, CINF 068.
49. James, C.A., Weininger, D., and Delany, J. (2002) *Daylight Theory Manual.* Daylight Chemical Information Systems, Inc.
50. Kanerva, P. (1990) *Sparse Distributed Memory.* MIT Press, Cambridge, Massachusetts, pp. 26–27.
51. Klir, G.J. and Yuan, B. (1995) *Fuzzy Sets and Fuzzy Logic: Theory and Applications.* Prentice Hall PTR, Upper Saddle River, New Jersey.
52. Miyamoto, S. (1990) *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Kluwer Academic Publishers, Dordrecht, The Netherlands.
53. Maggiora, G.M., Petke, J.D., and Mestres, J. (2002) A general analysis of field-based molecular similarity indices. *J. Math. Chem.* **31**, 251–270.
54. Hurst, T. and Heritage, T. (1997) HQSAR – A highly predictive QSAR technique based on molecular holograms. 213th ACS Natl. Meeting, San Francisco, CA, CINF 019.
55. Schneider, G., Neidhart, W., Giller, T., and Schmid, G. (1999) "Scaffold-hopping" by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed.* **38**, 2894–2896.
56. Xue, L., Godden, J.W., and Bajorath, J. (1999) Database searching for compounds with similar biological activity using short binary bit string representations of molecules. *J. Chem. Inf. Comput. Sci.* **39**, 881–886.
57. Wikipedia website, http://en.wikipedia.org/wiki/Euclidean_vector (*Last accessed October 22, 2009*).
58. Hyvarinen, A., Karhunen, J., and Oja, E. (2001) *Independent Component Analysis.* Wiley, New York.
59. Kay, D.C. (1988) *Theory and Problems of Tensor Calculus, Schaum's Outline Series.* McGraw-Hill, New York.
60. Hodgkin, E.E. and Richards, W.G. (1987) Molecular similarity based on electrostatic potential and electric fields. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **14**, 105–110.
61. Good, A.C. and Richards, W.G. (1998) Explicit Calculation of 3D molecular similarity. *Perspec. Drug Disc. Design* **9/10/11**, 321–338.
62. Lemmen, C. and Lengauer, T. (2000) Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Design* **14**, 215–232.
63. Güner, O.F., Ed. (2000) *Pharmacophore Perception, Development and Use in Drug Design.* International University Line, La Jolla.
64. Mansfield, M.L., Covell, D.G., and Jernigan, R.L. (2002) A new class of molecular shape descriptors. Theory and properties. *J. Chem. Inf. Comput. Sci.* **42**, 259–273.
65. Grant, J.A., Gallardo, G.A., and Pickup, J.T. (1996) A fast method of molecular shape comparison. A simple application of a Gaussian description of molecular shape. *J. Comp. Chem.* **17**, 1653–1666.
66. Blinn, J.R., Rohrer, D.C., and Maggiora, G.M. (1998) Field-based similarity forcing in energy minimization and molecular matching. In *Pacific Symposium on Biocomputing '99*, R.B. Altman, *et al.*, Eds., World Scientific, Singapore, pp. 415–424.
67. Labute, P. (1999) Flexible alignment of small molecules. *J. Chem. Comput. Group*,

Spring 1999 Edition [http://www.chemcomp.com/feature/malign.htm].
68. Christoffersen, R.E. and Maggiora, G.M. (1969) *Ab initio* calculations on large molecules using molecular fragments. Preliminary investigations. *Chem. Phys. Letts.* **3**, 419–423.
69. Szabo, A. and Ostlund, N.S. (1982) *Modern Quantum Chemistry – Introduction to Advanced Electronic Structure Theory*. Macmillan Publishing Company, New York.
70. Kearsley, S.K. and Smith, G.M. (1990) An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Meth.* **3**, 615–633.
71. Lemmen, C., Hiller, C., and Lengauer, T. (1998) RigFit: A new approach to superimposing ligand molecules. *J. Comput.-Aided Mol. Design* **12**, 491–502.
72. Good, A.C., Hodgkin, E.E., and Richards, W.G. (1992) Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **32**, 188–191.
73. Carbó, R. and Calabuig, B. (1990) Molecular similarity and quantum chemistry. In *Concepts and Applications of Molecular Similarity*, M.A. Johnson and G.M. Maggiora, Eds.,Wiley-Interscience, New York, pp. 147–171.
74. Petitjean, M. (1995) Geometric molecular similarity from volume based distance minimization: Application to Saxitoxin and Tetrodotoxin. *J. Comput. Chem.* **16**, 80–90.
75. Petitjean, M. (1996) Three-dimensional pattern recognition from molecular distance minimization. *J. Chem. Inf. Comput. Sci.* **36**, 1038–1049.
76. Ballester, P.J. and Richards, W.G. (2007) Ultrafast shape recognition for similarity search in molecular databases. *Proc. Roy. Soc.* **A463**, 1307–1321.
77. Nissink, J.W.M., Verdonk, M.L., Kroon, J., Mietzner, T., and Klebe, G. (1997) Superposition of molecules: Electron density fitting by application of Fourier transforms. *J. Comput. Chem.* **18**, 638–645.
78. Keseru, G.M. and Kolossvary, I. (1999) *Molecular Mechanics and Conformational Analysis in Drug Design*. Wiley-Interscience (Blackwell Publishing), New York.
79. Jorgensen, W.L. and Tirado-Rives, J. (2005) Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6665–6670.
80. Lee, M.S., Salsbury, F.R., and Olson, M.A. (2004). An efficient hybrid explicit/implicit solvent method for biomolecular simulations. *J. Comput. Chem.* **25**, 1967–1978.
81. Chipot, C. and Pohorille, A., Eds. (2007) *Free Energy Calculations. Theory and Applications in Chemistry and Biology*. Springer, New York.
82. Petit, J., Meurice, N. and Maggiora, G.M. (2009) On the development of a "soft" Rule of Five. *J. Chem. Inf. Model.*, submitted.
83. Stephens, M. A. (1974) EDF Statistics for goodness of fit and some comparisons. *J. Am. Stat. Assoc.* **69**, 730–737.
84. Krishnan, V. (2006) *Probability and Random Processes*. Wiley-Interscience, Hoboken, New Jersey.
85. Martin, Y.C. (2001) Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem.* **3**, 231–250.
86. Seilo, G. (1998) Similarity measures: Is it possible to compare dissimilar structures? *J. Chem. Inf. Comput. Sci.* **38**, 691–701.
87. Medina-Franco, J.L., Martínez-Mayorga, K., Giulianotti. M.A., Houghten, R.A., and Pinilla, C. (2008) Visualization of chemical space in drug discovery. *Curr. Comput.-Aided Drug Design* **4,** 322–333.
88. Oprea, T.I. and Gottfries, J. (2001) Chemography: The art of navigating in chemical space. *J. Comb. Chem.*, **3,** 157–166.
89. Wishart, D.S.; Knox, C.; Guo, A.C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; and Woolsey, J. DrugBank: A comprehensive resource for *in silico* drug discovery and exploration, *Nucl. Acids Res.* **2006**, *34*, D668-D672. (http://www.drugbank.ca/databases. *Accessed July 6, 2009*)
90. Austin, C.P., Brady, L.S., Insel, T.R., and Collins, F.S. (2004) Molecular biology: NIH Molecular libraries initiative. *Science* **306**, 1138–1139. This library is freely accessible by querying 'MLSMR'in PubChem (http://pubchem.ncbi.nlm.nih.gov. *Accessed October 29, 2009*)
91. Patterson, D.E., Cramer, R.D., Ferguson, A. M., Clark, R.D., and Weinberger, L.E. (1996) Neighborhood behavior: A useful concept for validation of molecular diversity. *J. Med. Chem.* **39**, 3049–3059.
92. Bellman, R.E. (1961) *Adaptive Control Processes*. Princeton University Press, Princeton, New Jersey.
93. Hastie, T., Tibshirani, R., and Friedman, J. (2001) *The Elements of Statistical Learning*. Springer, New York.
94. Bishop, C. (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.
95. Raghavendra, A.S. and Maggiora, G.M. (2007) Molecular basis sets – A general similarity-based approach for representing

chemical spaces. *J. Chem. Info. Model.* **47**, 1328–1340.
96. Simovici, D.A. and Djeraba, C. (2008) *Mathematical Tools for Data Mining: Set Theory, Partial Orders, Combinatorics.* Springer, London, UK.
97. Lee, J.A. and Verleysen, M. (2007) *Nonlinear Dimensionality Reduction.* Springer, New York.
98. Walker, P.D., Maggiora, G.M., Johnson, M. A., Petke, J.D., and Mezey, P.G. (1995) Shape group-analysis of molecular similarity - Shape similarity of 6-membered aromatic ring-systems. *J. Chem. Inf. Comput. Sci.* **35**, 568–578.
99. Rarey, M. and Dixon, J.S. (1998) Feature trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Design* **12**, 471–490.
100. Agrafiotis, D.K. and Lobanov, V.S. (2000) Nonlinear mapping networks. *J. Chem. Inf. Comput. Sci.* **40**, 1356–1362.
101. Rassokhin, D., Lobanov, V.S. and Agrafiotis, D.K. (2000) Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. *J. Comput. Chem.* **21**, 1–14.
102. Xie, D., Tropsha, A., and Schlick, T. (2000) An efficient projection protocol for chemical databases: Singular value decomposition combined with truncated-Newton minimization. *J. Chem. Inf. Comput. Sci.* **40**, 167–177.
103. Kruskal, J. (1977) The relationship between multidimensional scaling and clustering in *Classification and Clustering.* J. Van Ryzin, Ed., Academic Press, New York.
104. Gower, J.C. (1966) Some distance properties of latent roots and vector methods used in multivariate analysis. *Biometrika* **53**, 325–338.
105. Diamantaras, K.I. and Kung, S.Y. (1996) *Principal component neural networks – Theory and Applications.* Wiley, New York.
106. Benigni, R. and Giuliani, A. Analysis of distance matrices for studying data structures and separating classes. *Struct.-Act. Relat.* **12**, 397–401.
107. Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857–74.
108. Gower, J.C. (1984) Distance matrices and their Euclidean approximation. In *Data Analysis and Informatics, III*, E. Diday *et al.*, Eds., Elsevier Science Publishers B.V. (North-Holland).
109. Gower, J.C. and Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classific.* **3**, 5–48.
110. Benigni, R. (1994) EVE, a distance-based approach for discriminating non-linearly separable groups. *Quant. Struct.-Act. Relat.* **13**, 406–411.
111. Tenenbaum, J.B., de Silva, V., and Langford, J.V. (2000) A global geometric framework for non-linear dimensionality reduction. *Science* **290**, 2319–2323.
112. Roweis, S.T. and Saul, L.K. (2000) Nonlinear dimensionality reduction by local linear embedding. *Science* **290**, 2323–2326.
113. Friedman, J. and Tukey, J. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.* **C23**, 881–889.
114. Agrafiotis, D.K. (2003) Stochastic proximity embedding. *J. Comput. Chem.* **24**, 1215–1221.
115. Agrafiotis, D.K. and Xu, H. (2003) A geodesic framework for analyzing molecular similarities. *J. Chem. Inf. Comput. Sci.* **43**, 475–484.
116. Donoho, D.L. and Grimes, C. (2003) Hessian eigenmaps: Local linear embedding techniques for high-dimensional data. *Proc. Natl. Acad. Sci U. S. A.* **100**, 5591–55.
117. Maggiora, G.M., Shanmugasundaram, V., Lajiness, M.S., Doman, T.N., and Schulz, M.W. (2005) A practical strategy for directed compound acquisition. In *Chemoinformatics in Drug Discovery*, T.I. Oprea, Ed., pp. 317–332.
118. Maggiora, G.M. (2006) On outliers and activity cliffs – Why QSAR often disappoints. *J. Chem. Inf. Model.* **46**, 1535 (Editorial).
119. Doweyko, A.M. (2008) QSAR: dead or alive? *J. Comput.-Aided Mol. Design* **22**, 81–89.
120. Johnson, S. (2008) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* **48**, 25–26.
121. Guha, R. and Van Drie, J.H. (2008) Assessing how well a modeling protocol capture a structure-activity landscape. *J. Chem. Inf. Model.* **48**, 1716–1728.
122. Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M.S., and Van Drie, J.H. (2009) Navigating structure-activity landscapes. *Drug Disc. Today* **14**, 698–705.
123. Shanmugasundaram, V. and Maggiora, G.M. (2001) Characterizing property and activity landscapes using an information-theoretic approach. *222nd American Chemical Society Meeting*, Division of Chemical Information Abstract no. 77.
124. Renner, S. and Schneider, G. (2005) Scaffold-hopping potential of ligand-based similarity concepts. *Chem. Med. Chem.* **1**, 181–185.

125. Schneider, G., Schneider, P., and Renner, S. (2006) Scaffold hopping: How far can you jump? *QSAR Combin. Sci.* **25**, 1162–1171.
126. Maggiora, G.M. and Shanmugasundaram, V. (2005) An information-theoretic characterization of partitioned property spaces. *J. Math. Chem.* **38,** 1–20.
127. Medina-Franco, J.L., Maggiora, G.M., Giulianotti, M.A., Pinilla, C., and Houghten, R.A. (2007) A similarity-based data-fusion approach to the visual characterization and comparison of compound databases. *Chem. Biol. Drug Design* **70**, 393–412.
128. Guha, R. and Van Drie, J.H. (2008) Structure-activity landscape index: Identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **48**, 646–658.
129. Peltason, L. and Bajorath, J. (2007) SAR index: Quantifying the nature of structure-activity relationships. J. Med. Chem. **50**, 5571–5578.
130. Wawer, M., Peltason, L., Weskamp, N., Teckentrup, A., and Bajorath, J. (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J. Med. Chem.* **51**, 6075–6084.
131. Medina-Franco, J.L., Martínez-Mayorga, K., Bender, A., Marín, R.M., Giulianotti, M.A., Pinilla, C., and Houghten, R.A. (2009) Characterization of activity landscapes using 2D and 3D similarity methods: *Consensus activity cliffs. J. Chem. Inf. Model.* **49**, 477–491.
132. Christoffersen, R.E. (1989) *Basic Principles and Techniques of Molecular Quantum Mechanics.* Springer, New York.
133. Schölkopf, B. and Smola, A. (2002) *Learning with Kernels.* MIT Press, Cambridge, MA.
134. Herbrich, R. (2002) *Learning Kernel Classifiers.* MIT Press, Cambridge, MA.
135. Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis.* Cambridge University Press, Cambridge, UK.
136. Löwdin, P.O. (1992) On linear algebra, the least square method, and the search for linear relations by regression analysis in quantum chemistry and other sciences. *Adv. Quantum Chem.* **23**, 83–126.
137. Meyer, C.D. (2000) *Matrix Analysis and Applied Linear Algebra.* Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.
138. Carlson, B.C. and Keller, J.M. (1957) Orthogonalization procedures and the localization of Wannier functions. *Phys. Rev.* **105**, 102–103.
139. Agrafiotis, D.K., Rassokhin, D.N., and Lobanov, V.S. (2001) Multi-dimensional scaling and visualization of large molecular similarity tables. *J. Comput. Chem.* **22**, 1–13.
140. Kauvar, L.M., Higgins, D.L., and Villar, H.O., *et al.* (1995) Predicting ligand binding to proteins by affinity fingerprinting. *Chem. Biol.* **2**, 107–118.
141. Randic, M. (1991) Resolution of ambiguities in structure-property studies by use of orthogonalized descriptors. *J. Chem. Inf. Comput. Sci.* **31**, 311–320.
142. Randic, M. (1991) Correlation of enthalpy of octanes with orthogonal connectivity indices. *J. Mol. Struct.(Theochem)* **233**, 45–59.
143. Randic, M. (1993) Fitting non-linear regressions by orthogonalized power series. *J. Comput. Chem.* **14**, 363–370.

http://www.springer.com/978-1-60761-838-6

Chemoinformatics and Computational Chemical Biology
Bajorath, J. (Ed.)
2011, X, 588 p. 150 illus., 13 illus. in color., Hardcover
ISBN: 978-1-60761-838-6
A product of Humana Press