Chapter 1
Introduction to Molecular Similarity and Chemical Space

Gerald M. Maggiora

List of Abbreviations

2-D	Two-dimensional
3-D	Three-dimensional
APFs	Atom pair fingerprints
CS	Chemical space
CSN	Chemical space network
DB	Database
ECFPs	Extended connectivity fingerprints
FPs	Fingerprints
HOMO	Highest-occupied molecular orbital
HTS	High-throughput screening
LBVS	Ligand-based virtual screening
LUMO	Lowest-unoccupied molecular orbital
MaxD	Maximum dissimilarity selection (‘Dfragall’) algorithm
MaxST	Maximum spanning tree
MinST	Minimum spanning tree
MDS	Multidimensional scaling
NLM	Nonlinear mapping
PC	Principal component
PCA	Principal component analysis
PCoA	Principal coordinate analysis
PSA	Post-search aggregation
P-S	Pearlman–Smith

G. M. Maggiora (✉)
University of Arizona BIO5 Institute, 1657 East Helen Street, Tucson, AZ 85721, USA
e-mail: gerry.maggiora@gmail.com

Translational Genomics Research Institute, 445 North Fifth Street, Phoenix, AZ 85004, USA

K. Martinez-Mayorga, J. L. Medina-Franco (eds.), *Foodinformatics,*
DOI 10.1007/978-3-319-10226-9_1

1.1 Introduction

It is estimated that the chemical universe associated with small organic molecules is nearly 200 billion [1]. An older estimate, which includes larger organic molecules up to a molecular weight of 500 Da, suggests that this number may be around 10^{60} [2] and constitutes what could be called the "small molecule universe." Enumerating and searching this set of compounds would be a daunting task. Recently, a new approach has been published that is based on the construction of what the authors claim is a "representative universal library" of drug-like compounds [3]. In any case, regardless of how the size of the chemical universe is assessed, there is no question that its size is immense. Because of the size of even "representative" subsets of that universe, computer-based methods are required to capture, manage, and search the massive amount of information, activities that fall under the rubric of chemical informatics.

While the chemical universe of molecules potentially relevant in food science is considerably smaller, it nonetheless is large enough to benefit from many of the chemical informatic concepts that have proved useful in medicinal chemistry and related fields of chemistry. Two of these concepts, molecular similarity and chemical space (CS), are dealt with in this chapter. Of the two, molecular similarity is more fundamental since it plays a crucial role in the definition of CS itself. Though important, activity or property landscapes, which provide the third leg of a triad of activities that play important roles in much of chemical informatics, will not be discussed here. Numerous recent publications describing the visual and statistical aspects of activity landscapes as well as the basic features of these landscapes should be consulted for details [4–8].

Similarity is a ubiquitous concept that touches nearly every aspect of our conscious lives and, no doubt, influences our subconscious thoughts as well. Although its earliest influence on scientific thinking can be traced to the Greek philosophers [9, 10], its impact in chemistry began in the nineteenth century, the most notable example being the development of the periodic table of elements by Mendeleev [11] and Meyer [12]. As noted by Rouvray (see Table 1.4 in [9]), the twentieth century saw a significant expansion in the number and variety of chemical applications of molecular similarity. However, it was not until late in that century that application of similarity flourished due in large measure to the greater availability of digital computers. This led to the development of a plethora of methods for computing molecular similarity, enabling medicinal chemists to address a growing need to search compound collections[1] of rapidly increasing size for molecules with similar properties or biological activities.

Underlying this effort was the *similarity-property principle* (SPP) [13–15], which simply states that "Similar molecules tend to have similar properties." Although perhaps intuitively obvious, it nonetheless provides an important rationale that has proved quite helpful as a basis for similarity searches of CSs.

[1] The term database (DB) will generally be used to describe large collection of compounds whether or not material exists for screening the compounds.

However, because similarity is a subjective concept ("Similarity like pornography is difficult to define, but you know it when you see it" [10]), an absolute standard to judge the effectiveness of similarity methods does not exist. As will be discussed in the sequel, this raises some significant issues that can seriously impact the effectiveness and reliability of similarity methods; chief among them is the fact that the similarity values depend on the method used to encode the relevant chemical or molecular information. Nevertheless, a large number of successful applications have shown that similarity methods, with all of their inherent flaws, can provide an effective means for carrying out a number of chemical informatic activities that facilitate the practice of medicinal chemistry and drug discovery (*vide infra*). There are two main approaches to similarity in chemistry, what is typically called *molecular* or *structural similarity*, which is the focus of this work, and *chemical similarity*. The chemical similarity typically, but not exclusively, utilizes representations associated with macroscopic chemical properties such as solubility, heat of vaporization, molar refractivity, and log*P*, although occasionally properties of individual molecules such as pi-electron densities, highest-occupied and lowest-unoccupied molecular orbital (HOMO and LUMO) energies, and dipole moments are also used.

Representations associated with molecular similarity are in general classified as one-dimensional (1-D), two-dimensional (2-D), or three-dimensional (3-D). 1-D representations generally refer to macroscopic (e.g., solubility, log*P,* sublimation energy, heat of formation, etc.) or microscopic (e.g., molecular orbital energy, atomic charges, spectra, etc.) scalar quantities (*vide supra*). 2-D features are derived from the 2-D structures typically used by chemists to represent molecules. Although such structures can encode stereochemical and conformational information, this is not generally the case in molecular similarity studies, which typically use what are called hydrogen-suppressed chemical graphs [16], where hydrogen atoms, except those on specific nitrogen and oxygen atoms, are not explicitly represented. Thus, chemical graphs primarily encode information on the types of atoms and the bonds between them—the latter is sometimes referred to as the *bond topology* of the molecule.

By contrast, 3-D features are generally derived from the overall 3-D geometric, and sometimes the electronic structure of molecules, which would seem to provide a more faithful representation of molecular information. Nevertheless, a number of substantive issues remain. This is especially true of molecules with multiple conformational states, since determining what conformational state or states have to be included in a given similarity analysis is not entirely straightforward. For example, in similarity studies aimed at identifying molecules with comparable biological activities to known active molecules, does one use the minimum energy conformation or the biologically active one, which in many cases is not known. What about the case, when there are multiple conformations of comparable energy? All of these issues can significantly complicate 3-D similarity studies.

Because of the greater simplicity of 2-D compared to 3-D representations, and because the corresponding functions used to evaluate similarities are generally easier to carry out as well, 2-D similarities tend to be much faster to compute than the

3-D similarities (see Sect. 1.2 for details). This raises the question of whether 2-D similarities perform equally or better than 3-D methods in tasks commonly carried out in chemical informatics. Conclusive results have not been achieved to date. Nevertheless, it appears that 2-D methods can in many cases perform equally well and in some cases outperform 3-D methods [17, 18] in a variety of tasks. These tasks include similarity-based searches designed to identify new, potentially active molecules based on previously determined actives and to identify molecules with potentially similar values for properties of interest in drug research such as log*P*—both are examples of the SPP. In addition, these workers showed that of the 2-D methods considered "molecular ACCess system" (MACCS) structural-key-based fingerprints (FPs) (*vide infra*) consistently exhibited the best performance.

Because of this, most applications of molecular similarity over large sets of compounds generally employ 2-D similarity methods. It should be emphasized, however, that procedures for comparing 2-D versus 3-D similarity methods are imperfect by their very nature since, as noted earlier, similarity is a subjective concept that does not admit to absolute comparisons of any type.

In simplistic terms, the concept of CS can be considered to be a multidimensional extension of the concept of a congeneric series. However, an important distinction between the two is that CS involves a *pairwise relation* that specifies the relationship of the molecules to each other, generally in terms of a molecular similarity or CS-distance function. A set of objects and a pairwise relation among them are the basic ingredients of a mathematical space. In the present case, the objects are molecules and the pairwise relation characterizes the similarity or distance of separation of each pair of molecules in the CS. Similarity and distance are inversely related; the more similar a pair of molecules, the closer they are in CS, and vice versa.

Because CSs are generally of high dimension, faithfully depicting them in 2-D or 3-D is not possible, and some type of approximation is required. This, however, is not generally a problem because their visual depiction is only used *qualitatively*. More *quantitative* results can be obtained simply by carrying out the computations with respect to the full dimension of the CS in question.

Importantly, CS provides a conceptual framework for organizing the structural and property relationships of vast numbers of molecules within a common framework. With the burgeoning amount of structural, chemical, and biological data currently being created and stored in publically accessible databases (DBs) such as ChEMBL [19], PubChem [20], ChemDB [21], and DrugBank [22], or in subscription-based DBs such as WOMBAT [23] and MDDR [24], a conceptual framework, such as that provided by CS, is essential if we are to gain insights from information stored in these DBs. A summary of many public and private compound DBs is given in [25].

The remainder of this chapter covers set- and vector-based representations of structural and molecular data and how this information is converted into the various similarity, dissimilarity, and distance measures that have found wide application in chemical informatics. Examples of some of the types of structural and molecular descriptors are also presented, along with a discussion of their essential features. Significant emphasis is given to the concept of CS, a concept that plays an absolutely essential role in almost all aspects of chemical informatics. Finally, examples of how similarity can be used to carry out many activities associated with CSs, such as comparing compound collections, acquiring new compounds to augment current collections, assessing the diversity of a collection, generating diverse subsets of compounds for high-throughput screening (HTS) campaigns, and ligand-based virtual screening (LBVS). The latter activity has risen in importance over the past decade as an important strategy in drug discovery. The words "molecule" and "compound," which are very similar and are quite prevalent throughout this work, are used essentially interchangeably.

Over the past decade, a number of books have provided a good overview of many aspects of the field of chemical informatics [26–30], and a number of reviews and papers on molecular similarity [31–34] and CS [35–40] have also been published. These sources should be consulted for additional details on any of the subjects discussed in this work.

This chapter is not meant as a comprehensive review of molecular similarity and CSs. Rather it is intended to be somewhat pedagogical and to present, in some detail, a number of their key features and the interrelationships among them. In this way, it is hoped that readers will have a basic feel for the nature of the concepts and will be able to move on from there to tackle more complex aspects of these concepts and to apply them in a practical setting.

1.2 Structural Similarity Measures

Structural similarity is a pairwise relation between molecules. Similarity values are determined by a *similarity measure* that has three key components: (1) a representation of the relevant chemical or structural features of the molecules being compared, (2) an appropriate weighting of these features, and (3) a function that maps the feature information for pairs of molecules to a value that lies on the unit interval of the real line [0,1]. As noted in the previous section, representations can utilize macroscopic chemical features, electronic structural features of individual molecules, and/or geometric features associated with the structure or substructures of molecules

A number of procedures for computing 2-D and 3-D molecular similarities have been described in great detail [10, 41]. In the current work, the focus is on the class of 2-D similarity measures based on molecular FPs that encode the substructural information in molecules and on measures derived from vectors whose components represent macroscopic and microscopic physicochemical properties or indices derived from the topological properties of their chemical graphs. These approaches are the most prevalent ones and have been applied in a wide range of applications cited in Sects. 1.1, 1.2.3, 1.3.1, 1.3.2.1, 1.3.5.2, and 1.3.5.3. Moreover, they provide clear examples of the general workings of the types of molecular similarity measures in wide use today.

1.2.1 Set-Based Similarity Measures

1.2.1.1 Set-Based Representations: Binary Structural FPs

Consider the set of n molecules

$$\mathrm{M} = \left\{ \mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_i, \ldots, \mathrm{M}_n \right\}. \tag{1.1}$$

A *binary molecular FP* for molecule M_i can be specified by a set of p substructural features

$$\mathbf{m}_i = \left\{ m_i(1), m_i(2), \ldots, m_i(k), \ldots, m_i(p) \right\} \tag{1.2}$$

where the binary values of the indicator (characteristic) functions $m_i(k), k = 1, 2, \ldots, p$ in Eq. (1.2) determine whether a specific substructural feature is present or absent in the molecule, i.e.,

$$m_i(k) = \begin{cases} 1, \text{ if the } k\text{th structural feature is present;} \\ 0, \text{ if the } k\text{th structural feature is absent.} \end{cases} \tag{1.3}$$

Binary molecular FPs are sometimes called *bit vectors* or *bit strings* since their elements are “1s” and “0s”. In this work, the nomenclature *binary molecular FP* may also be given by *structural FP, molecular FP, binary FP*, or just *FP*. Multiple occurrences of structural features are not accounted for in binary FPs, although they can be as described later in this section.

Equation (1.4) depicts a hypothetical FP

$$\mathbf{m}_i = \overbrace{(1, 0, 0, 1, 1, \ldots, 1, 0)}^{p} \tag{1.4}$$

characterized by a *binary p*-tuple. This is a reasonably standard notation for FPs. However, because of their sparseness (i.e., relatively few 1-bits), it is not how they are generally handled in computers, where *index-based* and *run-length* encoding schemes are typically used [42]. The former scheme basically indexes all of the 1-bits of a given FP. By contrast, run-length encoding indexes the lengths of runs of 0-bits followed by a 1-bit. As an illustration, consider the following simple example of a binary structural FP (0,0,1,1,0,0,0,0,1,0,0,1,0,0,0). Hence, its index-based encoding is given by (2,0,4,2), while its corresponding run-length encoding is (3,4,9,12), an example that clearly shows that the encodings are not of fixed length. Except in a few instances where stereochemical information is represented, most molecular FPs are based on the 2-D structural features of the molecules they symbolize and, hence, the representations described in this work correspond to 2-D molecular FPs. The number of components or elements, p, in molecular FPs can be quite large and can be either fixed or variable. The former usually corresponds to *molecule-independent* FPs and the latter to *molecule-dependent* FPs.

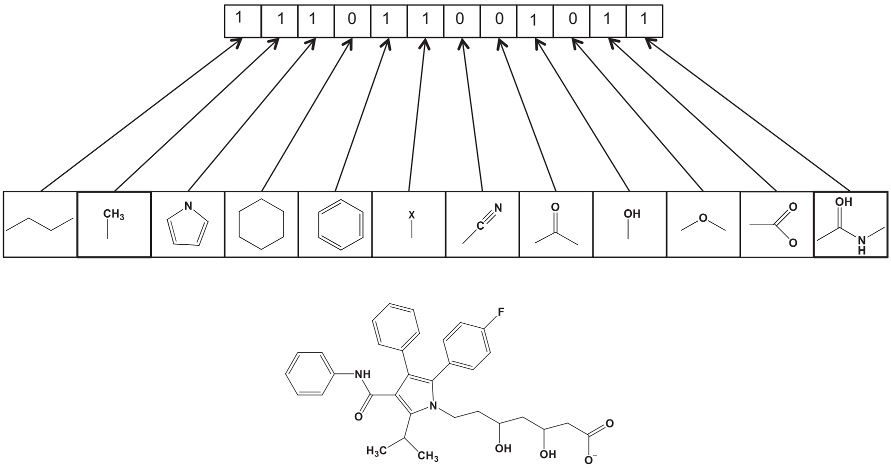


Fig. 1.1 An example based on the drug Lipitor of a simplified molecule-independent directory-based binary structural FP with its corresponding set of descriptors. The symbol 'X' corresponds to any of the halogen atoms (F, Cl, Br, I)

Molecule-Independent/Directory-Based FPs The number of structural features in molecule-independent FPs is fixed for all molecules, as exemplified by MACCS key FPs, which contain 166 structural features [43] and Barnard Chemical Information (BCI) FPs that contain more than 1000 features [44]. Figure 1.1 provides a simple example, based on the anticholesterol drug Lipitor, of a molecule-independent FP. Note that multiple occurrences of methyl groups, hydroxyl groups, and phenyl rings are not explicitly accounted for, nor is the elongated hydrocarbon chain that connects the nitrogen atom of the pyrrole ring with the terminal carboxylate fully accounted for, although a structural descriptor that represents a shorter hydrocarbon chain provides at least a partial account of the elongated chain.

Hence, structural information can be lost leading to similarity values of unity for pairs of molecules that are not structurally identical. Nevertheless, there is at least a partial correspondence between the descriptors in the directory and the binary molecular FP of a molecule, so that it may be possible in many instances to associate particular substructural features with molecular properties and/or biological activities, a characteristic that is not generally shared by molecule-dependent FP representations (*vide infra*). This can be partially ameliorated through the use of *weighted molecular FPs* that take account of the number of times a structural feature occurs in a molecule. However, since not all structural features that may be associated with a specific structure–property relationship (SPR) or structure–activity relationship (SAR) are necessarily accounted for in given FP, it may not be possible to infer SPR or SAR even when weighted FPs are employed.

Molecule-dependent FPs have variable numbers of elements that typically depend on the number of non-hydrogen atoms and functional complexity of molecules. Because of the rapid growth in the size and molecular complexity of modern compound DBs, molecule-dependent FPs have been growing in popularity since they can potentially handle a wider range of molecules than molecule-independent FPs. Two structural FPs that exemplify the types of molecule-dependent FPs in use today are the atom pair FPs (APFs) first developed by Carhart, Smith, and Venkataraghaven nearly 30 years ago [45] and the more recent extended connectivity FPs (ECFPs) developed by Rogers and Hahn [46] that are in widespread use today. Simple examples of APFs and ECFPs are depicted in Figs. 1.2 and 1.3**,** respectively.

Both of these FPs are referred to as “2-D FPs,” since neither of them utilizes 3-D structural information. Although a number of FPs including AFPs and ECFPs can encode stereochemical information, they rarely do in common usage.

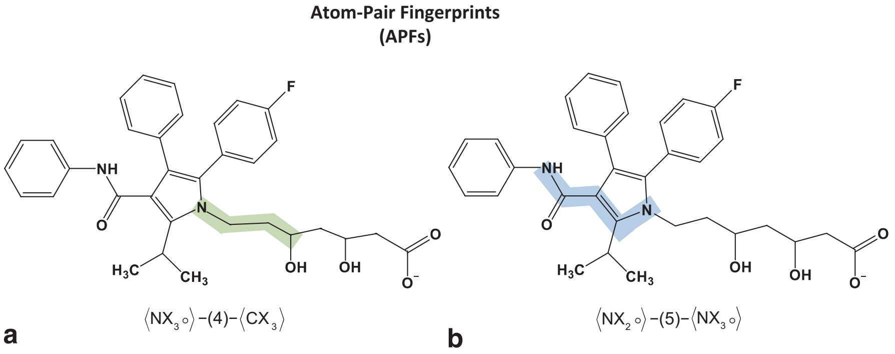


Fig. 1.2 Examples of molecule-dependent atom pair fingerprints (APF) descriptors depicted with respect to the drug Lipitor. Regions highlighted in *light green* and *light blue* correspond to substructures associated with two APFs; the labels below each figure correspond to respective designations given in reference [46] for these APFs

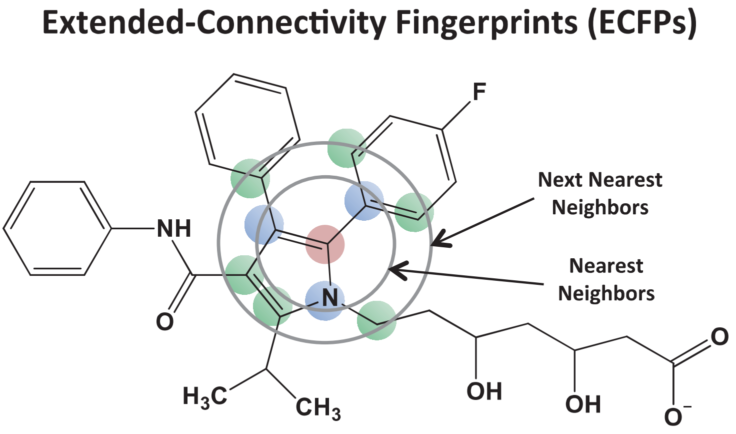


Fig. 1.3 Examples of molecule-dependent extended connectivity ECFP descriptors depicted with respect to the drug Lipitor. Atoms lying within the rings depicted in the figure correspond to nearest (*colored in light blue*) and next-nearest neighbors (*colored in light green*) to the central atom (*colored in light red*) of a given ECFP4 descriptor.

Atom Pair FPs Pairs of atoms and the minimum number of bonds linking them constitute the substructural components of APFs. Generally, only APs separated by seven or fewer bonds are considered. As described by Carhardt et al. [45], the general form of the substructure of an APF is given by Eq. (1.5):

$$\left\langle \text{“atom-}i\text{”} \right\rangle - \left(\text{atom-separation}\right) - \left\langle \text{“atom-}j\text{”} \right\rangle, \tag{1.5}$$

where “atom-*i*” and “atom-*j*” are descriptions that contain information on the atom type (e.g., C, N, O,…), the number of non-hydrogen atoms bound to it, and whether it possesses a bonding pi-electron. The “separation” between atoms is based on a count of all the atoms, including atom-*i* and atom-*j*, on the shortest through-bond path connecting the two terminal atoms of the chain. Consider, for example, the APF designation $\left\langle \mathrm{NX}_3 \bullet \right\rangle - (4) - \left\langle \mathrm{CX}_3 \right\rangle$ depicted in Fig. 1.2a. In the $\mathrm{NX}_3 \bullet$ term contained within the leftmost brackets, “N” designates the leftmost atom in the chain highlighted in light green, “X_3” indicates that three atoms are bonded to it, and the “•” indicates the presence of bonding pi-electron on the nitrogen atom. Next, the “4” in the parentheses indicates the number of atoms in the chain including the terminal atoms. Last, in the CX_3 term contained within the rightmost brackets, “C” designates the rightmost atom in the chain and “X_3” indicates that three atoms are bonded to it. A similar interpretation can be made for the designation corresponding to the APF highlighted in light blue in Fig. 1.2b.

Because of the way in which APFs are handled in a computer, it is not possible to associate substructural features with specific bits in an APF. An excellent discussion based on the closely related Daylight FPs [47] discusses this issue and many other of the technical details that must be addressed in order to effectively implement APFs.

Extended Connectivity FPs By contrast, ECFPs sample the molecular environment surrounding each non-hydrogen atom. Thus, the local “circular” environments surrounding each non-hydrogen atom constitute the substructural features of a given molecule as depicted in Fig. 1.3. Although not always employed, ECFPs can also encode stereochemical information, which can be important in many aspects of drug discovery research since all stereoisomers of a given compound may not be equally active.

For example, consider the pyrrolic carbon atom in Fig. 1.3 highlighted in light red. As seen in the figure, two layers of atoms surround it, the first, whose atoms are highlighted in light blue, corresponds to nearest neighbors and the second, whose atoms are highlighted in light green, corresponds to next nearest neighbors. Each non-hydrogen atom and its layers of surrounding atoms constitute substructural features. The maximum number of layers considered is given by the diameter of the largest circular environment surrounding the central atom. This is based on the number of bonds needed to connect two diametrically opposed atoms in that layer. In the case shown here, four bonds are required. Such FPs are designated by ECFP4.

From the above, it is easy to see that the number of possible FP descriptors that can be obtained for compound collections is quite large. For example, Rogers and

Hahn [46] have shown that sets of ~50,000 compounds can give rise to ECFP descriptors that number in the hundreds of thousands. For larger sets of compounds, the number of ECFP descriptors can potentially exceed 1 million. Hence, handling this amount of information efficiently presents some technical problems, the details of which are beyond the scope of this work. Interestingly, unlike AFPs whose substructural information cannot be retrieved, this is not the case for ECFPs, although the procedure for doing so requires several steps. The paper by Rogers and Hahn [46] provides a detailed discussion of many of these issues. They also note that ECFPs were designed primarily to characterize the activities of compounds. Hence, ECFPs contain information on features that are present as well as those that are not present. ChemAxon provides a very clear description of many of the technical details associated with application of ECFPs [48]. In addition, they offer a useful, albeit brief, comparative discussion of AFPs and ECFPs, pointing out that the former performs best for substructure searches while the latter appears to be more suitable for similarity searches. Several other papers also provide useful assessments of ECFPs [49, 50].

Weighted Structural FPs Weighting the features of structural FPs is not common practice in chemical informatics. Nevertheless, it has been shown in a number of studies to provide improved results in virtual screening experiments [51–53].

Although numerous schemes exist [54], weighting nowadays is typically accomplished by accounting in some fashion for the number of occurrences of each of the features in a molecule, as for example, the methyl, phenyl, hydroxyl groups depicted in Fig. 1.1 for the hypercholesterol drug Lipitor™.

Clearly, not accounting for multiple occurrences of features can lead to significant *degeneracies* that arise when different compounds have identical FPs. Sometimes the degeneracies can be quite large as shown by the following analysis based on Lipitor™. Consider each of the multiple structural FP descriptors in Lipitor™: three phenyl, two methyl, and two hydroxyl groups. There are seven possible descriptor patterns containing at least one phenyl group and three possible patterns containing at least one methyl group and three containing at least one hydroxyl group. Assuming that each of the three descriptor patterns is independent of each other, a quite reasonable assumption is that the total number of possible patterns is $7 \times 3 \times 3 = 67$. Hence, there are 67 different, albeit related, compounds that would all have exactly the same structural FP as Lipitor™. While this may be a somewhat extreme example, there are nonetheless numerous examples of compounds with multiple occurrences of specific substructural patterns. Surprisingly, the results obtained with unweighted FPs are quite good. And although both APFs and ECFPs can take account of multiple occurrences of substructural patterns, they are rarely if ever considered in actual applications.

In fact, most cheminformatic studies continue to use binary structural FPs.

1.2.1.2 FP-Based Similarity Coefficients

The third component of a similarity measure is the function that maps the structural information contained in the molecular FPs of each pair of compounds

Table 1.1 Set-theoretic expressions useful in molecular similarity analysis

Symbol	Set-theoretic expression[a]	Definition
N_i	$\mathrm{Card}(\mathbf{m}_i)$	Number of features in molecule M_i
$N_{i,j}$	$\mathrm{Card}(\mathbf{m}_i \cap \mathbf{m}_j)$	Number of features common to molecules M_i and M_j
$N_i - N_{i,j}$	$\mathrm{Card}(\mathbf{m}_i) - \mathrm{Card}(\mathbf{m}_i \cap \mathbf{m}_j)$	Number of features unique to molecule M_i

[a] "Card" refers to the cardinality (i.e., number of elements) of the set in question

being compared to the unit interval of the real line [0,1]. Such functions are called by a number of names—*similarity functions, similarity indices*, or *similarity coefficients*—the latter nomenclature will be adhered to in this chapter [10]. Although there are many types of similarity coefficients, only a limited number will be considered here. A summary of all types of similarity coefficients is given in a comprehensive review [31].

Based on his work in mathematical psychology, Tversky developed the most general form of similarity coefficient applicable to structural FPs [55]:

$$S_{\mathrm{Tve}}(i,j|\alpha,\beta) = \frac{N_{i,j}}{\alpha(N_i - N_{i,j}) + \beta(N_j - N_{i,j}) + N_{i,j}}, \tag{1.6}$$

where the weighting parameters satisfy $\alpha,\beta \geq 0$, which ensures that the similarity values lie on the unit interval of the real line [0,1]. The various terms in Eq. (1.6) are described in Table 1.1.

As described in Table 1.1, the terms in parentheses in the denominator, $N_i - N_{i,j}$ and, $N_j - N_{i,j}$, can be interpreted as the number of features unique to molecules M_i and M_j, respectively, weighted by the corresponding values of α and β.

It is clear from the form of Eq. (1.6) that the Tversky similarity coefficient is generally *asymmetric* with respect to the interchange of its arguments, i.e., $\mathrm{M}_i \rightarrow \mathrm{M}_j$ and $\mathrm{M}_i \leftarrow \mathrm{M}_j$. This corresponds to interchanging the associated variables N_i and N_j in Eq. (1.6) so that $(N_i \rightarrow N_j$ and $N_i \leftarrow N_j)$, i.e.,

$$S_{\mathrm{Tve}}(j,i|\alpha,\beta) = \frac{N_{i,j}}{\alpha(N_j - N_{i,j}) + \beta(N_i - N_{i,j}) + N_{i,j}} \tag{1.7}$$

which is equal to the expression in Eq. (1.6) and is symmetric *only* in cases where $\alpha = \beta$: Note that the variable $N_{i,j}$ is invariant to these interchanges. Such cases correspond to well-known similarity coefficients, three of which are described below.

For example, the currently most popular similarity coefficient, $S_{\mathrm{Tan}}(i,j)$, is that due to Tanimoto and is obtained by setting $\alpha = \beta = 1$,

$$S_{\mathrm{Tve}}(i,j|\alpha=1,\beta=1) = \frac{N_{i,j}}{(N_i - N_{i,j}) + (N_j - N_{i,j}) + N_{i,j}} = S_{\mathrm{Tan}}(i,j). \tag{1.8}$$

The sum of the terms in the denominator is equal to the total number of features in common plus the number of unique features associated with molecules M_i and M_j, although the form of the expression differs from that usually used, namely, $N_i + N_j - N_{i,j}$, where the "$-N_{i,j}$" term corrects for double counting the features in both molecules. Thus, the Tanimoto similarity coefficient is the ratio of the number of features in common to both molecules over the total number of features (not the sum) in M_i and M_j.

Setting $\alpha = \beta = \frac{1}{2}$ leads to the *Dice similarity coefficient*:

$$S_{\mathrm{Tve}}(i, j \,|\, \alpha = \frac{1}{2}, \beta = \frac{1}{2}) = \frac{N_{i,j}}{\frac{1}{2}(N_i + N_j)} = S_{\mathrm{Dice}}(i, j), \tag{1.9}$$

where the term in the denominator is the *arithmetic mean* of the number of features in M_i and M_j. Thus, the Dice similarity coefficient is the ratio of the number of features in common to M_i and M_j over the arithmetic mean of the number of their features.

Although it cannot be obtained from $S_{\mathrm{Tve}}(i, j \,|\, \alpha, \beta)$ simply by choosing appropriate values for α and β, the well-known *cosine similarity coefficient* given by

$$S_{\cos}(i, j) = \frac{N_{i,j}}{\sqrt{N_i \cdot N_j}} = \frac{N_{i,j}}{N_i^{\frac{1}{2}} \cdot N_j^{\frac{1}{2}}} \tag{1.10}$$

can be obtained from a related but more general similarity function [56]. Interestingly, the denominator is the *geometric mean* of the number of elements in M_i and M_j, so that the cosine similarity coefficient is the ratio of the number of features in common to M_i and M_j over the geometric mean of the features.

Although not as general as the expression given in Eq. (1.6), a useful expression is obtained by setting $\beta = 1 - \alpha$, which gives

$$S_{\mathrm{Tve}}(i, j \,|\, \alpha) = \frac{N_{i,j}}{\alpha(N_i - N_{i,j}) + (1-\alpha)\cdot(N_j - N_{i,j}) + N_{i,j}} \tag{1.11}$$

so that $\alpha + \beta = 1$. Under such a constraint, it is not possible to transform Eq. (1.6) into the expression for Tanimoto similarity, Eq. (1.8), although the Dice coefficient given in Eq. (1.9) can still be obtained by setting $\alpha = 1/2$. Any value of $\alpha \neq 1/2$ leads to asymmetric similarity coefficients. This asymmetry has been applied to enhance the effectiveness of similarity searches of large compound DBs [57, 58].

An interesting pair of asymmetric similarity coefficients is obtained at the limits when $\alpha = 1$ or $\alpha = 0$:

$$S_{\mathrm{Tve}}(i, j \,|\, \alpha = 1) = \frac{N_{i,j}}{(N_i - N_{i,j}) + N_{i,j}} = \frac{N_{i,j}}{N_i} \tag{1.12}$$

and

$$S_{\mathrm{Tve}}(i,j|\alpha=0)=\frac{N_{i,j}}{(N_j-N_{i,j})+N_{i,j}}=\frac{N_{i,j}}{N_j}. \tag{1.13}$$

Equation (1.12) can be interpreted as the fraction of M_i similar to M_j, while Eq. (1.13) can be interpreted as the fraction of M_j similar to M_i. By applying the “interchange rules” to Eq. (1.12), it is clear that the similarity coefficients are *asymmetric*, i.e.,

$$\begin{aligned}S_{\mathrm{Tve}}(i,j|\alpha=1)&=\frac{N_{i,j}}{(N_j-N_{i,j})+N_{i,j}}=\frac{N_{i,j}}{N_j}\neq\frac{N_{i,j}}{N_i}.\\&=S_{\mathrm{Tve}}(j,i|\alpha=1)\end{aligned} \tag{1.14}$$

A similar argument can be applied to Eq. (1.13).

Symmetric similarity coefficients corresponding to the asymmetric coefficients are given in Eqs. (1.15) and (1.16) and can be obtained simply by changing the denominators using the “min” and “max” functions, which are symmetric to interchanges of variables N_i and N_j:

$$S_{\mathrm{Min}}(i,j)=\frac{N_{i,j}}{\min(N_i,N_j)} \tag{1.15}$$

and

$$S_{\mathrm{Max}}(i,j)=\frac{N_{i,j}}{\max(N_i,N_j)}. \tag{1.16}$$

As was the case for the other similarity coefficients, S_{Max} and S_{Min} are again ratios equal to the number of features common to M_i and $\mathrm{M}_j$over the larger and smaller number of features of M_i and M_j, respectively.

It can be shown that all of the similarity coefficients described above lie on the unit interval [0,1]. Because the terms in the denominators satisfy the following inequalities:

$$0<\min(N_i,N_j)\le N_i^{\frac{1}{2}}\cdot N_j^{\frac{1}{2}}\le\tfrac{1}{2}(N_i+N_j)\le\max(N_i,N_j)\le N_i+N_j-N_{i,j} \tag{1.17}$$

and because their numerators are all identical and equal to $N_{i,j}$, the five symmetric similarity coefficients are ordered as:

$$0<S_{\mathrm{Tan}}\le S_{\mathrm{Max}}\le S_{\mathrm{Dice}}\le S_{\mathrm{Cos}}\le S_{\mathrm{Min}}\le 1. \tag{1.18}$$

1.2.1.3 FP-Based Molecular Dissimilarity Coefficients

For FP-based representations, dissimilarity is the 1's *complement* of similarity, i.e.,

$$\text{Dissimilarity} = 1 - \text{similarity}. \tag{1.19}$$

Thus, dissimilarity values also lie on the unit interval [0,1]. For example, in the case of the Tanimoto similarity coefficient the corresponding dissimilarity coefficient is given by

$$D_{\text{Tan}}(i, j) = 1 - S_{\text{Tan}}(i, j) \tag{1.20}$$

which is symmetric because $S_{\text{Tan}}(i, j)$ is symmetric. Substituting Eq. (1.8) into Eq. (1.20) and simplifying terms yields

$$D_{\text{Tan}}(i, j) = \frac{(N_i - N_{i,j}) + (N_j - N_{i,j})}{(N_i - N_{i,j}) + (N_j - N_{i,j}) + N_{i,j}}. \tag{1.21}$$

Since the denominators, which normalize the similarity and dissimilarity values, in Eqs. (1.8) and (1.21), respectively, are the same for both coefficients, it is their numerators that provide the interpretation for these coefficients. In the case of Tanimoto similarity, the numerator, $N_{i,j}$, gives the number of features in common to both molecules, while the numerator for Tanimoto dissimilarity gives the number of features unique to M_i, $N_i - N_{i,j}$, and the number of features unique to M_j, $N_j - N_{i,j}$. This interpretation accords well with our qualitative notions of similarity and dissimilarity. *Features that do not appear in either molecule are not accounted for in any of these coefficients*.

It can also be shown that Tanimoto dissimilarity formally satisfies the three properties of an abstract distance [59]. In fact, the numerator is identical to the Hamming distance between two finite, classical sets [60] and the denominator ensures that the dissimilarity values satisfy $0 \le D_{\text{Tan}} \le 1$, as required by Eq. (1.20).

Based on Eq. (1.19), dissimilarity coefficients corresponding to the similarity coefficients given in Eqs. (1.9), (1.10), (1.15), and (1.16) can also be constructed. Interestingly, the terms in their denominators are unchanged from their corresponding similarity coefficients. However, the terms in their numerators are the same as those in their denominators with the important difference that $N_i \to N_i - N_{i,j}$ and $N_j \to N_j - N_{i,j}$. Thus, for example, the Dice dissimilarity coefficient becomes

$$D_{\text{Dice}}(i, j) = \frac{\frac{1}{2}\left[(N_i - N_{i,j}) + (N_j - N_{i,j})\right]}{\frac{1}{2}(N_i + N_j)} \tag{1.22}$$

which is the ratio of the arithmetic mean of the number of unique features in M_i and M_j to the arithmetic mean of the total number of features in M_i and M_j. Recall that the term in square brackets is the Hamming distance so, as was the case for Tanimoto dissimilarity, Dice dissimilarity also satisfies the distance postulates.

Analogous expressions for dissimilarity can be derived for the remaining similarity coefficients.

1.2.1.4 Size Dependence of FP-Based Similarity and Dissimilarity Coefficients

It is both intuitive and well known that the number of 1-bits in a binary molecular FP depends on the size and complexity of the molecule it is representing. More than 25 years ago, Flower noted a bias towards low similarity values in Tanimoto similarity-based searches when the *bit densities*, that is the ratio of 1-bits to the total number of bits in a binary FP, of the molecules being compared differed significantly [61]. Subsequently, a number of laboratories observed a bias in diversity analyses towards smaller compounds [31, 62–65]. A publication also in that period by Godden et al. [66] further elaborated the issue by showing that mean Tanimoto similarity values obtained from sets of compounds are inherently biased by statistically preferred similarity values.[2]

It is not difficult to see how molecular size may have a biasing effect on the Tanimoto coefficient given in Eqs. (1.8). Consider two molecules, a *query molecule*, M_{Q}, and a *retrieved molecule*, M_{R}, obtained from a similarity search. Now suppose that the query molecule is a small molecule such that the number of substructural features (1-bits) in the FPs of both molecules satisfies $N_{\mathrm{Q}} < N_{\mathrm{R}}$. Since the number of substructural features common to both molecules, $N_{\mathrm{Q,R}}$, cannot be more than the number in the smaller of the two molecules,[3] i.e.,

$$N_{\mathrm{Q,R}} \le \max(N_{\mathrm{Q,R}}) = N_{\mathrm{Q}}. \tag{1.23}$$

In which case,

$$S_{\mathrm{Tan}}(Q,R) = \frac{N_{\mathrm{Q,R}}}{(N_{\mathrm{Q}} - N_{\mathrm{Q,R}}) + N_{\mathrm{R}}} \le \frac{\max(N_{\mathrm{Q,R}})}{\left[N_{\mathrm{Q}} - \max(N_{\mathrm{Q,R}})\right] + N_{\mathrm{R}}} = \frac{N_{\mathrm{Q}}}{N_{\mathrm{R}}} \tag{1.24}$$

The inequality obtains from Eq. (1.23) and the fact that the denominator of Eq. (1.24) satisfies

$$(N_{\mathrm{Q}} - N_{\mathrm{Q,R}}) + N_{\mathrm{R}} \ge N_{\mathrm{R}} \tag{1.25}$$

[2] Interestingly, since FP-based similarity coefficients are ratios of two integers, they represent a limited subset of rational numbers. Hence, they can by their very nature only yield restricted set of values on the unit interval of the real line.

[3] In that case, the set of features in M_Q are a subset of those in M_R.

Thus, for fixed values of N_{Q} and N_{R}, $S_{\mathrm{Tan}}(Q,R)$ reaches its maximum when the features of the query molecule are a subset of those of the retrieved molecule, that is, when $N_{\mathrm{Q,R}} \to \max(N_{\mathrm{Q,R}}) = N_{\mathrm{Q}}$. In this case, the smaller (or the closer in size) the retrieved molecule is to the query molecule, the larger the Tanimoto similarity value, and hence, the bias for small molecules in Tanimoto similarity searches when the query molecule is itself a small molecule. This type of bias should be called *algebraic bias* since it arises out of the form of the Tanimoto similarity coefficient and has no statistical component (cf. [67]).

If, on the other hand, the query is now a large molecule such that $N_{\mathrm{Q}} > N_{\mathrm{R}}$, then Eq. (1.26) can be obtained from Eq. (1.24) simply by interchanging the subscripts *Q* and *R*, i.e.,

$$S_{\mathrm{Tan}}(Q,R) = \frac{N_{\mathrm{Q,R}}}{(N_{\mathrm{R}} - N_{\mathrm{Q,R}}) + N_{\mathrm{Q}}} \leq \frac{\max(N_{\mathrm{Q,R}})}{\left[N_{\mathrm{R}} - \max(N_{\mathrm{Q,R}})\right] + N_{\mathrm{Q}}} = \frac{N_{\mathrm{R}}}{N_{\mathrm{Q}}} \quad (1.26)$$

It is clear from the equation that since the query molecule is large and fixed, the only way to increase $S_{\mathrm{Tan}}(Q,R)$ is to increase the size of the retrieved molecule. Hence, in Tanimoto similarity searches where the query molecule is large, something that rarely occurs in practice, the algebraic bias will be towards larger retrieved molecules. Holliday et al. [67] have significantly extended this analysis, providing an extensive and detailed treatment of a large number of similarity coefficients that are documented in Table 1.1 of their paper.

The algebraic bias in similarity searches has led some researchers to consider other possible similarity functions that might overcome this problem. An interesting work in this regard is that of Chen and Brown [57], which was based on asymmetric similarity searching. A detailed discussion of asymmetric similarity searching and how it might overcome, to some extent at least, the algebraic size bias described above was recently presented [10, 41].

Although the algebraic size bias discussed above is relatively straight forward, this is not case when dealing with dissimilarity-based searching as it is applied, for example, in diversity analysis. In each step of a typical iterative dissimilarity-based selection algorithm, the most dissimilar compound with respect to *all* of the previously selected compounds is chosen, a situation that differs significantly from that of similarity searching in a number of ways (see discussion in Sect. 1.3.5.3 for additional details). Moreover, the arguments presented above do not touch on some of the crucial issues that are statistical in nature. These were clearly described in a paper by Fligner et al. [65] and involved a statistical analysis of the discrete, hypercubical space in which binary structural FPs reside. Based on this analysis, they developed a modified version of the Tanimoto similarity coefficient that in addition to accounting for substructural features present in both molecules, also considered features that were absent. Basically, it is a weighted combination of Tanimoto similarity coefficients, one corresponding to the usual form of the Tanimoto coefficient associated with 1-bits and the other of essentially similar form but associated in this case with the 0-bits.

It was shown by both Fligner et al. [65] and Holliday et al. [67] that the modified Tanimoto coefficient did to a large extent ameliorate size bias associated with the Tanimoto similarity coefficient. More recently, Bajorath and his collaborators [58, 66] successfully introduced a related type of modified similarity measure that weights contributions associated with the presence and absence of substructural features. In their case, however, a Tverksy-type similarity coefficient was used rather than the Tanimoto expression employed by Fligner et al. [65].

1.2.2 Vector-Based Similarity Measures

Analogous expressions to the FP-based Tanimoto, Dice, and Cosine similarity coefficients (see Eqs. (1.8), (1.9), and (1.10), respectively) also exist for vectors with continuous, real valued components as described in the following section.[4] Since each of the vector components may be associated with properties that have different units, i.e., are not comparable, they can be *standardized* according to Eqs. (1.30) and (1.31), so that their values are mean centered and of unit variance. *Also, subscripts designating the similarity coefficient are given in bold face upper case type to distinguish them from the corresponding FP-based similarity coefficients*. Terms typically found in vector-based similarity and dissimilarity coefficients are described in Table 1.2.

Table 1.2 Vector-based expressions useful in similarity analysis

Operation	Vector expression	Corresponding set theoretic entities[a]
Scalar product of $\mathbf{x}_{\text{row}}(i)$ and $\mathbf{x}_{\text{row}}(j)$	$\langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j) \rangle = \sum_{k=1}^{p} x_{ik} x_{jk}$	N_i
Squared magnitude of $\mathbf{x}_{\text{row}}(i)$	$\lVert \mathbf{x}_{\text{row}}(i) \rVert^2 = \langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(i) \rangle = \sum_{k=1}^{p} x_{ik}^2$	$N_{i,j}$
Squared Euclidean distance between $\mathbf{x}_{\text{row}}(i)$ and $\mathbf{x}_{\text{row}}(j)$	$\lVert \mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j) \rVert^2$ $= \langle \mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j), \mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j) \rangle$ $= \sum_{k=1}^{p} (x_{\text{row}}(i) - x_{\text{row}}(j))^2$	$(N_i - N_{i,j})$ $+ (N_j - N_{i,j})$

[a] See Table 1.1

[4] Strictly speaking, these vectors should be called geometric vectors since they do not, in all cases, satisfy the properties of algebraic vectors (e.g., algebraic vectors satisfy the axioms of a linear vector space, namely, the addition of two vectors or the multiplication of a vector by a scalar should result in another vector that also lies in the space). Nevertheless, the terminology "vector," which is common in chemical informatics, will be used here to include both classes of vectors.

1.2.2.1 Vector-Based Representations

Vector-based representations provide another means for encoding the molecular and chemical information associated with molecule M_i and are of the general form of *p*-dimensional row vectors also called *p*-tuples:

$$\mathbf{x}_{\mathrm{row}}(i) = (x_{i,1}, x_{i,2}, \ldots, x_{i,k}, \ldots, x_{i,p}), \quad i = 1, 2, \ldots n \tag{1.27}$$

Such vectors are in many instances given as column vectors. However, since the rows of data matrices generally correspond to points in a data space, the practice is continued here for consistency.

Each component of the vector represents the value of some macroscopic chemical property such as solubility, heat capacity, polarizability, pK_a [68], some molecular property such as molecular weight, ionization potential, pi-electron distribution, number of hydrogen bonding donors or acceptors, and HOMO or LUMO energies [69], or some properties that characterize topological aspects of molecules, such as branching and shape indices [70]. Martin [71] has discussed the computation of many physicochemical property descriptors in the context of computational drug design. Todeschini and Consonni have compiled an extensive compendium of them [72]; Guha and Willighagen have recently surveyed a wide variety of quantitative descriptors useful for the calculation of chemical and biological properties [73]. Labute has also developed an internally consistent set of 32 descriptors based on the surface properties of molecules such as log*P*, molar refractivity, partial charges, and pK_as [74, 75]. They were shown to be weakly correlated with each other, able to represent much of the information in many "traditional" molecular descriptors, and capable of providing an effective means for carrying out a range of quantitative structure–activity relationship (QSAR) and structure–property relationship (QSPR) calculations.

BCUT Descriptors A particularly interesting set of descriptors is that developed by Pearlman and Smith [76–78]. Called BCUTS, they provide an internally consistent, balanced set of molecular descriptors that encode information on the electrostatic, hydrophobic, and hydrogen bonding features of molecules and are generated in a way that exploits information on through-bond or through-space interatomic distances and atomic properties related to intermolecular ligand–protein interactions. BCUT values are determined from matrices whose diagonal elements are associated with atomic properties and whose off-diagonal elements are associated with connectivity-related properties and a scale factor that balances both types of information. Different definitions of the off-diagonal elements differentiate the different classes of BCUTS from each other. For example, 3-D BCUTS use through space interatomic distances to determine off-diagonal elements, while 2-D BCUTS use Burden numbers [79], and 2-DT BCUTS use topological interatomic distances. The largest and smallest eigenvalue obtained from each matrix are retained as potential descriptors.

Since there are many ways to compute the diagonal and off-diagonal elements of BCUT matrices, the number of potential descriptors is quite large for any of the three BCUT classes. In order to deal with this issue, Pearlman and Smith developed an "auto-choose" algorithm based on a χ-squared statistic that selects an optimum subset of BCUT descriptors for a given set of compounds such that their distribution is as close to a uniform distribution as possible. Thus, intercompound correlations are reduced so that the compounds are maximally dispersed throughout CS in the minimum number of dimensions. Importantly, this shows that BCUT descriptors and their associated CSs depend on the set of compounds used to determine them. Thus, there are many possible CSs, most typically of dimension five and six. BCUT descriptor values are not standardized to zero mean and unit variance (see Eqs. 1.30 and 1.31) since their value ranges are all comparable.

BCUT descriptors have been shown to perform well in diversity-analysis-related tasks [80–82]. And although not originally intended for this purpose BCUT descriptors have, nonetheless, shown surprisingly good performance in QSAR and QSPR studies [83–85] and in selecting compounds for follow-on screening in drug discovery [86].

In general, the vectors associated with a set of n molecules can be combined into an $n \times p-$dimensional data matrix

$$\mathbf{X}_{nxp} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,j} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,j} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & x_{i,2} & \cdots & x_{i,j} & \cdots & x_{i,p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,j} & \cdots & x_{n,p} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{\text{row}}(1) \\ \mathbf{x}_{\text{row}}(2) \\ \vdots \\ \mathbf{x}_{\text{row}}(i) \\ \vdots \\ \mathbf{x}_{\text{row}}(n) \end{bmatrix} \tag{1.28}$$

where ith row is the same as that given by Eq. (1.27) and jth column is given by

$$\mathbf{x}_{\text{col}}(j) = \begin{pmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{k,j} \\ \vdots \\ x_{n,j} \end{pmatrix}, \quad j = 1, 2, \ldots, p \tag{1.29}$$

Because the units associated with each of the descriptors are, in general, likely to differ, they should be normalized so that they all have equivalent units. This can be accomplished by standardizing the set of values for each descriptor to zero mean and unit variance using the well-known "z-transformation," i.e.

$$z_{i,j} = \frac{x_{i,j} - \bar{x}_{\text{col}}(j)}{\sqrt{s(j)}} \tag{1.30}$$

where the sample mean and variance of the jth variable are given by, respectively,

$$\begin{aligned}\overline{x}_{\text{col}}(j) &= \tfrac{1}{n}\sum_{i=1}^{n} x_{i,j}\\ s(j) &= \tfrac{1}{n}\sum_{i=1}^{n}\left[x_{i,j} - \overline{x}_{\text{col}}(j)\right]^2\end{aligned} \tag{1.31}$$

All of the variables are now unitless and, thus, on equal footing. Row vectors and data matrices corresponding to the new z-transformed variables are now given, respectively, by (cf. Eqs. 1.30 and 1.31)

$$\mathbf{z}_{\text{row}}(i) = (z_{i,1}, z_{i,2}, \ldots, z_{i,k}, \ldots, z_{i,p}),\ \ i = 1, 2, \ldots . n \tag{1.32}$$

and

$$\mathbf{Z}_{nxp} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,j} & \cdots & z_{1,p} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,j} & \cdots & z_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{i,1} & z_{i,2} & \cdots & z_{i,j} & \cdots & z_{i,p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ z_{n,1} & z_{n,2} & \cdots & z_{n,j} & \cdots & z_{n,p} \end{bmatrix} = \begin{bmatrix} \mathbf{z}_{\text{row}}(1) \\ \mathbf{z}_{\text{row}}(2) \\ \vdots \\ \mathbf{z}_{\text{row}}(i) \\ \vdots \\ \mathbf{z}_{\text{row}}(n) \end{bmatrix} \tag{1.33}$$

1.2.2.2 Vector-Based Similarity Coefficients

The vector-based Tanimoto similarity coefficient corresponding to the FP-based coefficient in Eq. (1.8) is given by

$$\begin{aligned} S_{\mathbf{TAN}}(i,j) &= \frac{\left\langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j)\right\rangle}{\left\|\mathbf{x}_{\text{row}}(i)\right\|^2 + \left\|\mathbf{x}_{\text{row}}(j)\right\|^2 - \left\langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j)\right\rangle} \\ &= \frac{\sum_{k=1}^{p} x_{ik} \cdot x_{jk}}{\sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - \sum_{k=1}^{p} x_{ik} \cdot x_{jk}}, \end{aligned} \tag{1.34}$$

where the form of the continuous, real valued vectors is given in Eq. (1.27) and the nature of their components are described in the previous section. The vector-based similarity coefficient due to Hodgkin and Richards [87] is an analog of the FP-based Dice similarity coefficient given in Eq. (1.9):

$$\begin{aligned} S_{\mathbf{HR}}(i,j) &= \frac{\left\langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j)\right\rangle}{\tfrac{1}{2}(\left\|\mathbf{x}_{\text{row}}(i)\right\|^2 + \left\|\mathbf{x}_{\text{row}}(j)\right\|^2)} \\ &= \frac{\sum_{k=1}^{p} x_{ik} \cdot x_{jk}}{\tfrac{1}{2}\left(\sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2\right)} \end{aligned} \tag{1.35}$$

The well-known *cosine similarity coefficient*, also called the *Carbo similarity index* [10], provides a measure of the cosine of the angle between two vectors

$$\begin{aligned} S_{\mathbf{COS}}(i,j) &= \frac{\langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j) \rangle}{\sqrt{\|\mathbf{x}_{\text{row}}(i)\|^2 \cdot \|\mathbf{x}_{\text{row}}(j)\|^2}} \\ &= \frac{\sum_{k=1}^{p} x_{i,k} \cdot x_{j,k}}{\sqrt{\sum_{k=1}^{p} x_{i,k}^2} \cdot \sqrt{\sum_{k=1}^{p} x_{j,k}^2}} \end{aligned} \tag{1.36}$$

A variety of function and vector-based similarity coefficients have also been described [10], and a detailed analysis of their interrelationships has been presented [56].[5]

1.2.2.3 Vector-Based Dissimilarity Coefficients and Distances

Vector-based dissimilarity coefficients can also be defined in analogy to those given in general for FP-based dissimilarities in Eq. (1.19). Tanimoto dissimilarities are given by

$$\begin{aligned} D_{\mathbf{TAN}}(i,j) &= 1 - S_{\mathbf{TAN}}(i,j) \\ &= \frac{\langle \mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j), \mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j) \rangle}{\|\mathbf{x}_{\text{row}}(i)\|^2 + \|\mathbf{x}_{\text{row}}(j)\|^2 - \langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j) \rangle} \\ &= \frac{\|\mathbf{x}_{\text{row}}(i) - \mathbf{x}_{\text{row}}(j)\|^2}{\|\mathbf{x}_{\text{row}}(i)\|^2 + \|\mathbf{x}_{\text{row}}(j)\|^2 - \langle \mathbf{x}_{\text{row}}(i), \mathbf{x}_{\text{row}}(j) \rangle} \\ &= \frac{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}{\sum_{k=1}^{p} x_{ik}^2 + \sum_{k=1}^{p} x_{jk}^2 - \sum_{k=1}^{p} x_{ik} \cdot x_{jk}} \end{aligned} \tag{1.37}$$

Again, the terms are analogous to those for the FP-based dissimilarity given in Eq. (1.21) and summarized in Tables 1.1 and 1.2. As was the case for FP-based dissimilarities, the value of the vector-based dissimilarity is complementary (see Eq. 1.20) to the corresponding similarity value and, hence, lies on the unit interval

[5] An interesting relationship between the FP- and vector-based similarity coefficients occurs when both have binary component values, e.g. $\mathbf{m}_l = (1,0,0,0,1,1,0,1,0,1)$ and $\mathbf{x}_{\text{row}}(l) = (1,0,0,0,1,1,0,1,0,1)$. In such cases, but only in such cases, the similarity coefficients based on binary FPs or binary vectors yield exactly the same similarity value for all of the similarity coefficients described above. However, this limitation has not been consistently adhered to and similarity values computed using continuous vectors or weighted FPs based on Eqs. (1.27)–(1.29) yield values that may differ significantly from their corresponding FP-based similarity coefficients.

[0,1]. Importantly, the numerator is just the square of the *Euclidean distance* (see also Table 1.2):

$$
\begin{aligned}
d_{\mathrm{Euc}}(\mathbf{x}_{i\bullet},\mathbf{x}_{j\bullet})^2 &= \left\langle (\mathbf{x}_{i\bullet}-\mathbf{x}_{j\bullet}),(\mathbf{x}_{i\bullet}-\mathbf{x}_{j\bullet}) \right\rangle \\
&= \left\| \mathbf{x}_{i\bullet}-\mathbf{x}_{j\bullet} \right\|^2 \\
&= \sum_{k=1}^{p} (x_{ik}-x_{jk})^2
\end{aligned}
\tag{1.38}
$$

Since the denominator is just a constant factor that scales the distance so that distance lies on the unit interval [0,1], it again follows that $D_{\mathbf{TAN}}$ satisfies the distance axioms as was true in the corresponding FP-based case for D_{tan} .

Similarly, it can be shown that Hodgkin–Richards dissimilarity,

$$
D_{\mathbf{HR}}(i,j) = \frac{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}{\frac{1}{2}\left(\sum_{k=1}^{p}x_{ik}^2+\sum_{k=1}^{p}x_{jk}^2\right)}
\tag{1.39}
$$

accords well with the FP-based case for $D_{\mathrm{Dice}}(i,j)$. Note that the numerator is the squared Euclidean distance of the two molecular feature vectors, so dissimilarity again satisfies the distance axioms and is a normalized distance whose values lie on [0,1].

Thus, it is clear from the above discussion that there is an underlying consistency to the FP- and vector-based similarity coefficients. Moreover, for the case of binary FPs and binary feature vectors, the two approaches yield identical results (*vide supra*). However, for *integer-weighted FPs* (see Sect. 1.2.1.1) such as arise in cases where the number of occurrences of substructural features is considered, methods for treating vectors with continuous, real-valued components are no longer appropriate and multiset procedures provide a better, more consistent approach for dealing with such FPs [10, 41].

1.2.3 Fusing ("Aggregating") Similarity Measures

Although molecular similarity studies have been carried out for more than two decades, it is generally recognized that no one similarity measure is capable of providing high-quality results for all classes of compounds. This has raised the possibility that aggregating or fusing multiple similarity measures may in some fashion lead to improved results [88]. Based on the pioneering works of Sheridan and his colleagues at Merck [89, 90] and Peter Willett and his colleagues in Sheffield, a number of procedures have been developed for combining similarity measures based on data fusion methods [91–93]. A recent review by Willett provides a comprehensive overall summary and analysis of similarity-based data fusion methods [94].

Table 1.3 Examples of fusion rules

Fusion rule	Applicable fusion method[a]	Mathematical expression
MAX	Group fusion	$\max\left\{S_i^{\mathrm{Ref}_1}, S_i^{\mathrm{Ref}_2}, \ldots, S_i^{\mathrm{Ref}_q}\right\}$
MIN	Similarity fusion	$\min\left\{R_i^{\mathrm{Sim}_1}, R_i^{\mathrm{Sim}_2}, \ldots, R_i^{\mathrm{Sim}_p}\right\}$
MEAN	Similarity fusion	$(1/p)\sum_{k=1}^{p} S_i^{\mathrm{Sim}_k}$
RRF	Similarity and group fusion	$\sum_{k=1}^{p}(1/R_i^{\mathrm{Sim}_k})$ or $\sum_{l=1}^{q}(1/R_i^{\mathrm{Ref}_l})$

Only the most effective fusion rules are included in the table, where "S_i" corresponds to a similarity value and "R_i" to a specific rank. See text for details

[a] See [94] for a detailed discussion of the performance of the different fusion rules

Data fusion methods [95, 96] fall under the more general rubric of *data aggregation* methods that are widespread in many applications of multiparameter decision making [97]. The basic idea behind data fusion is that combining data from multiple sources will lead to improved results over data obtained from a single source. Data fusion can be implemented as an *unsupervised* or *supervised* procedure, the former being the most well studied of the two approaches, since the latter requires experimental activity data in addition to computed similarities [94]. The focus in this work is on unsupervised procedures, and the previous reference should be consulted for details of supervised procedures. The description of similarity searching given in Sect. 1.3.3.3 is complementary to that presented here, where the emphasis is on issues associated with data fusion procedures.

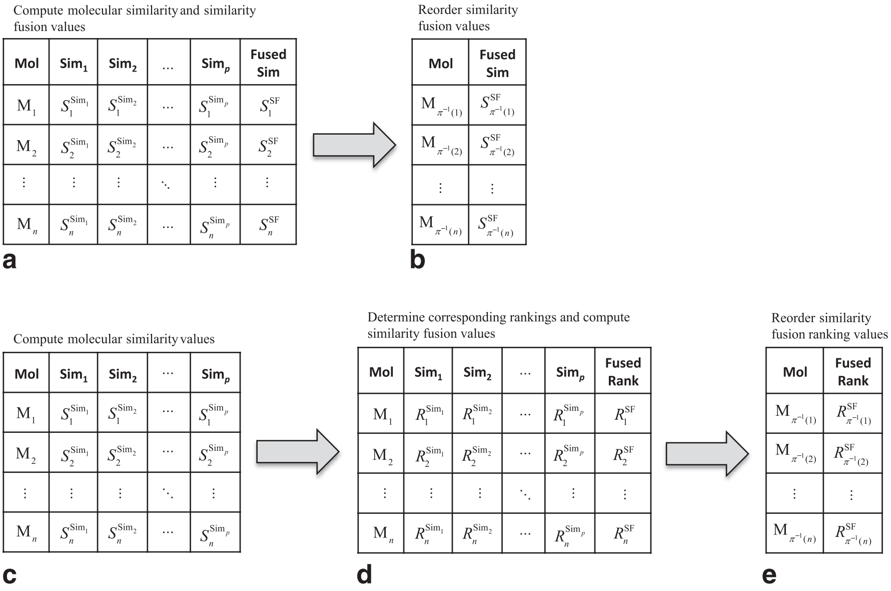
Although there are many possible unsupervised ways to combine multisource data, those typically applied in chemical informatics are relatively limited (see Fig. 1.2 of [94]). Table 1.3 provides a summary of the most effective data fusion rules associated with the different fusion procedures typically employed in chemical informatics applications (*vide infra*). In certain applications, as seen in the table, data are best treated as similarity values or as rankings—specifics are described below. Mathematical expressions corresponding to the different fusion rules given in Table 1.3 are relatively straightforward except for the reciprocal rank fusion (RRF) rule, which is directly related to the mean of the harmonic mean of the rank values [98].[6] Because the RRF rule treats rank values reciprocally, compounds near the top of a ranked list will have lower values, and thus will be given more influence in the RRF rule than those further down the list. Recent studies in information retrieval [99] and chemical informatics [100] suggest that the RRF rule may be more generally applicable than heretofore had been suspected. Thus, it may be suitable as a replacement for the other fusion rules considered in Table 1.3 (i.e., MAX, MIN, and MEAN), which have enjoyed widespread use in the past [94]. Finally, it should be noted that fusions can also be effected using similarity values computed with any of the similarity measures although most studies have been confined to FP-based measures.

[6] The RRF rule works best with rank values since similarities can in certain cases have zero values leading to undefined values for the reciprocals, a situation that can be overcome by the addition of a small positive constant to the denominator of each term.

1.2.3.1 Similarity Fusion

The initial approach to data fusion, called *similarity fusion*, combines the results of searches using multiple similarity measures with respect to a single reference molecule. The data generated in this procedure can be envisioned in the form of a data table such as that depicted in Fig. 1.4a, where the columns correspond to the p different similarity measures, and the rows correspond to the n molecules in a DB—at this point the ordering of the molecules is arbitrary. Each of the similarity elements in the table, $S_i^{\mathrm{Sim}_k}$, is designated by the DB molecule with which it is associated, as indicated by the set of subscripts $\{1, 2, \ldots, n\}$. The corresponding similarity measures used to calculate its value are indicated by the set of superscripts $\{\mathrm{Sim}_1, \mathrm{Sim}_2, \ldots, \mathrm{Sim}_p\}$. *All of the similarity values are computed with respect to the same reference molecule.*

The similarity values in each row can be aggregated in various ways to yield a fused similarity value S_i^{SF}. For example, as shown in Table 1.3, the arithmetic mean values of the similarity values in each row can be computed and placed in the corresponding column "fused sim" of Fig. 1.4a. Once this process is complete the rows can be reordered, as depicted in Fig. 1.4b, such that the first row contains the most similar molecule to the reference molecule based on its fused similarity value, the second row contains the next most similar molecule, and the process continues until all of the molecules are reordered with respect to their fused similarity values. This procedure effectively *permutes* the order of the molecules given in the first column of Fig. 1.4a, which as noted earlier is arbitrary, to that shown in the first column of



Compute molecular similarity and similarity fusion values

Mol	Sim$_1$	Sim$_2$	…	Sim$_p$	Fused Sim
M_1	$S_1^{\mathrm{Sim}_1}$	$S_1^{\mathrm{Sim}_2}$	…	$S_1^{\mathrm{Sim}_p}$	S_1^{SF}
M_2	$S_2^{\mathrm{Sim}_1}$	$S_2^{\mathrm{Sim}_2}$	…	$S_2^{\mathrm{Sim}_p}$	S_2^{SF}
⋮	⋮	⋮	⋱	⋮	⋮
M_n	$S_n^{\mathrm{Sim}_1}$	$S_n^{\mathrm{Sim}_2}$	…	$S_n^{\mathrm{Sim}_p}$	S_n^{SF}

a

Reorder similarity fusion values

Mol	Fused Sim
$M_{\pi^{-1}(1)}$	$S_{\pi^{-1}(1)}^{\mathrm{SF}}$
$M_{\pi^{-1}(2)}$	$S_{\pi^{-1}(2)}^{\mathrm{SF}}$
⋮	⋮
$M_{\pi^{-1}(n)}$	$S_{\pi^{-1}(n)}^{\mathrm{SF}}$

b

Compute molecular similarity values

Mol	Sim$_1$	Sim$_2$	…	Sim$_p$
M_1	$S_1^{\mathrm{Sim}_1}$	$S_1^{\mathrm{Sim}_2}$	…	$S_1^{\mathrm{Sim}_p}$
M_2	$S_2^{\mathrm{Sim}_1}$	$S_2^{\mathrm{Sim}_2}$	…	$S_2^{\mathrm{Sim}_p}$
⋮	⋮	⋮	⋱	⋮
M_n	$S_n^{\mathrm{Sim}_1}$	$S_n^{\mathrm{Sim}_2}$	…	$S_n^{\mathrm{Sim}_p}$

c

Determine corresponding rankings and compute similarity fusion values

Mol	Sim$_1$	Sim$_2$	…	Sim$_p$	Fused Rank
M_1	$R_1^{\mathrm{Sim}_1}$	$R_1^{\mathrm{Sim}_2}$	…	$R_1^{\mathrm{Sim}_p}$	R_1^{SF}
M_2	$R_2^{\mathrm{Sim}_1}$	$R_2^{\mathrm{Sim}_2}$	…	$R_2^{\mathrm{Sim}_p}$	R_2^{SF}
⋮	⋮	⋮	⋱	⋮	⋮
M_n	$R_n^{\mathrm{Sim}_1}$	$R_n^{\mathrm{Sim}_2}$	…	$R_n^{\mathrm{Sim}_p}$	R_n^{SF}

d

Reorder similarity fusion ranking values

Mol	Fused Rank
$M_{\pi^{-1}(1)}$	$R_{\pi^{-1}(1)}^{\mathrm{SF}}$
$M_{\pi^{-1}(2)}$	$R_{\pi^{-1}(2)}^{\mathrm{SF}}$
⋮	⋮
$M_{\pi^{-1}(n)}$	$R_{\pi^{-1}(n)}^{\mathrm{SF}}$

e

Fig. 1.4 Data tables illustrating *similarity fusion* of similarity and rank values: **a** and **b** depict the procedure for fusing similarity values. **c**, **d**, and **e** depict the corresponding procedure for fusing rank values

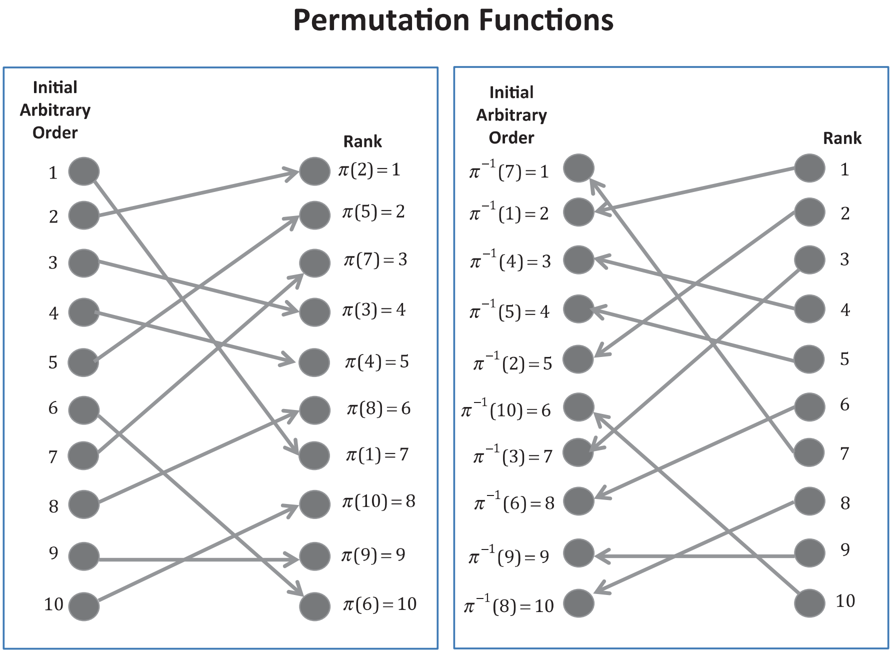


Fig. 1.5 Graphical example of mappings produced by the permutation functions π and their inverses π^{-1} (see text for additional details)

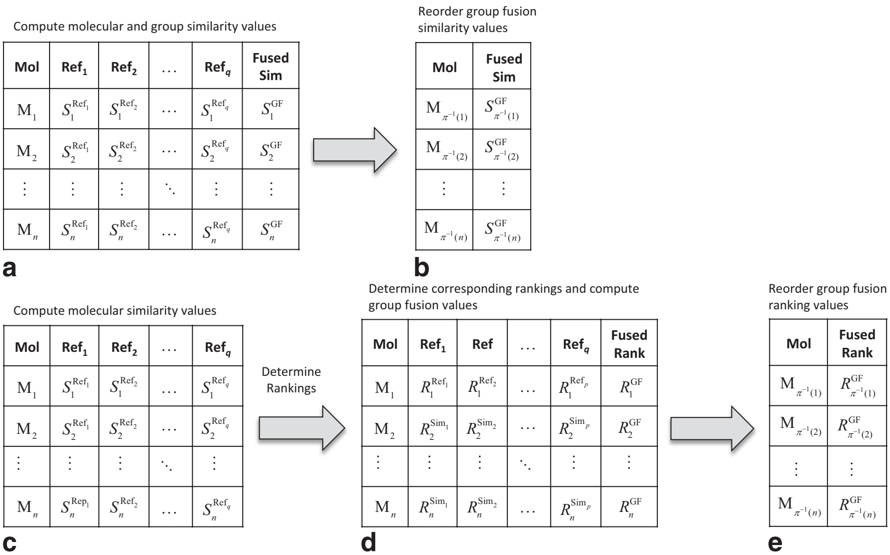
Fig. 1.4b, which is based on the decreasing fused similarity values in the second column of Fig. 1.4b, i.e.,

$$S^{\mathrm{SF}}_{\pi^{-1}(1)} \geq S^{\mathrm{SF}}_{\pi^{-1}(2)} \geq \cdots \geq S^{\mathrm{SF}}_{\pi^{-1}(n)} \tag{1.40}$$

The subscript notation $\pi^{-1}(i)$ in the mathematical expression given in Eq. (1.40) is based on the mathematical theory of permutations [101], where the permutation function value $\pi(i)$ gives the rank of the *i*th molecule and the unique inverse $\pi^{-1}(j)$ designates the *j*th molecule in the overall ranking. A graphic example of how these functions operate is provided in Fig. 1.5. It is important to note that while the permutations determine the rank order of the compounds, it is the similarity values themselves that are combined using the MEAN fusion rule in similarity fusion.

Alternatively, data fusion procedures can also be directly applied to rankings themselves as seen in Fig. 1.4c. In this case, the computation of similarities is followed by a determination of the rank of each of the compounds with respect to each of the similarity measures as illustrated in Fig. 1.4d, and an appropriate data fusion procedure, in this case the MIN rule given in Table 1.3 is applied. Lastly, the resulting MIN fused rankings are permuted, i.e., $R^{\mathrm{SF}}_{j} \to R^{\mathrm{SF}}_{\pi^{-1}(i)} = i$, in increasing order

$$R^{\mathrm{SF}}_{\pi^{-1}(1)} < R^{\mathrm{SF}}_{\pi^{-1}(2)} < \cdots < R^{\mathrm{SF}}_{\pi^{-1}(n)} \tag{1.41}$$



Compute molecular and group similarity values

Mol	Ref_1	Ref_2	…	Ref_q	Fused Sim
M_1	$S_1^{\text{Ref}_1}$	$S_1^{\text{Ref}_2}$	…	$S_1^{\text{Ref}_q}$	S_1^{GF}
M_2	$S_2^{\text{Ref}_1}$	$S_2^{\text{Ref}_2}$	…	$S_2^{\text{Ref}_q}$	S_2^{GF}
⋮	⋮	⋮	⋱	⋮	⋮
M_n	$S_n^{\text{Ref}_1}$	$S_n^{\text{Ref}_2}$	…	$S_n^{\text{Ref}_q}$	S_n^{GF}

Reorder group fusion similarity values

Mol	Fused Sim
$M_{\pi^{-1}(1)}$	$S_{\pi^{-1}(1)}^{\text{GF}}$
$M_{\pi^{-1}(2)}$	$S_{\pi^{-1}(2)}^{\text{GF}}$
⋮	⋮
$M_{\pi^{-1}(n)}$	$S_{\pi^{-1}(n)}^{\text{GF}}$

Compute molecular similarity values

Mol	Ref_1	Ref_2	…	Ref_q
M_1	$S_1^{\text{Ref}_1}$	$S_1^{\text{Ref}_2}$	…	$S_1^{\text{Ref}_q}$
M_2	$S_2^{\text{Ref}_1}$	$S_2^{\text{Ref}_2}$	…	$S_2^{\text{Ref}_q}$
⋮	⋮	⋮	⋱	⋮
M_n	$S_n^{\text{Rep}_1}$	$S_n^{\text{Ref}_2}$	…	$S_n^{\text{Ref}_q}$

Determine Rankings

Determine corresponding rankings and compute group fusion values

Mol	Ref_1	Ref	…	Ref_q	Fused Rank
M_1	$R_1^{\text{Ref}_1}$	$R_1^{\text{Ref}_2}$	…	$R_1^{\text{Ref}_p}$	R_1^{GF}
M_2	$R_2^{\text{Sim}_1}$	$R_2^{\text{Sim}_2}$	…	$R_2^{\text{Sim}_p}$	R_2^{GF}
⋮	⋮	⋮	⋱	⋮	⋮
M_n	$R_n^{\text{Sim}_1}$	$R_n^{\text{Sim}_2}$	…	$R_n^{\text{Sim}_p}$	R_n^{GF}

Reorder group fusion ranking values

Mol	Fused Rank
$M_{\pi^{-1}(1)}$	$R_{\pi^{-1}(1)}^{\text{GF}}$
$M_{\pi^{-1}(2)}$	$R_{\pi^{-1}(2)}^{\text{GF}}$
⋮	⋮
$M_{\pi^{-1}(n)}$	$R_{\pi^{-1}(n)}^{\text{GF}}$

Fig. 1.6 Data tables illustrating *group fusion* of similarity and rank values: **a** and **b** depict the procedure for group fusion of similarity values. **c**, **d**, and **e** depict the corresponding procedure for fusing rank values

1.2.3.2 Group Fusion

The development of *group fusion* [91, 92, 102] quickly followed that of similarity fusion. In contrast to latter, a *single similarity measure but multiple reference molecules* are used. This is illustrated in Fig. 1.6a, b, which are quite similar to the previous figure except that the similarity measures in the top row of Fig. 1.4a are replaced by a set of q reference molecules $\{\text{Ref}_1, \text{Ref}_2, \ldots, \text{Ref}_q\}$ in Fig. 1.6a. Similarity values are computed for each DB compound with respect to each of the reference compounds using a single similarity measure, and the values in each row of the table are fused, yielding the similarity values in the last column of Fig. 1.6a. As was the case for similarity fusion, the next step is to reorder the fused similarity values from largest to smallest as indicated in Fig. 1.6b and Eq. (1.42):

$$S_{\pi^{-1}(1)}^{\text{GF}} \geq S_{\pi^{-1}(2)}^{\text{GF}} \geq \cdots \geq S_{\pi^{-1}(n)}^{\text{GF}} \tag{1.42}$$

Numerous studies have shown that applying the MAX rule to similarity values provides excellent overall performance in similarity searches that are designed to assess the efficacy of group similarity for retrieving known actives from compound DBs [91, 92, 94, 100]. Although, in general, the MAX rule works well, the RRF rule for combining rank values (see Table 1.3) appears to perform even better [100]. Figure 1.6c–e describes the rank-based group fusion process, which is similar to that given in Fig. 1.4c–e for the corresponding similarity fusion procedure. The fused values obtained by the RRF rule are given in the far right column of Fig. 1.6d, and the combined values are then permuted, i.e., $R_j^{\mathrm{GF}} \to R_{\pi^{-1}(i)}^{\mathrm{GF}} = i$, in increasing order

$$R_{\pi^{-1}(1)}^{\mathrm{GF}} < R_{\pi^{-1}(2)}^{\mathrm{GF}} < \cdots < R_{\pi^{-1}(n)}^{\mathrm{GF}} \tag{1.43}$$

to yield the final fusion-based ranking. Whichever rule is used, the superior performance of group fusion makes it the preferred method for carrying out similarity searches [103].

Either the reordered similarity values or compound rankings can be used as a basis for subset selection. Furthermore, although group fusion provides improved results over both single similarity and similarity fusion approaches, it requires multiple reference compounds, which may not always be readily available. Even when such data are available, they usually are the result of early-phase HTS experiments and hence may, to a degree, be suspect. However, as discussed in the following section, a modification of group fusion called turbo similarity suggests that even somewhat erroneous data may not unduly affect the results obtained using group fusion.

1.2.3.3 Turbo Similarity

As noted in the previous section, a variant of group fusion called *turbo similarity* has also shown promise [104–106]. Turbo similarity provides a procedure for applying group fusion when only a single active is known and is based on the following procedures: (1) compute the similarity of the known (reference) active with respect to all of the molecules in a DB of unscreened compounds; (2) order the list with respect to decreasing similarity or increasing rank values; (3) choose a subset of the highest scoring or ranked compounds that, based on the SPP [13–15] (see Sect. 1.3.1 for details), are *assumed* to be active; and (4) use these putative active compounds as the set of reference compounds in a group-fusion-based similarity search as described in the previous section (see also Table 1.3 and Fig. 1.6). Note that either the MAX rule with respect to similarity or the RRF rule with respect to rank values can be applied with nearly comparable effectiveness (*vide supra*). A recent study [52] has shown that frequency weighting the components of structural FPs leads to improved results obtained with turbo similarity searching.

Interestingly, turbo similarity is reminiscent of library search procedures, where a given query yields a set of hits, each of which is used in a subsequent query to broaden the search [107].

1.2.4 Validating Similarity-Based Approaches

Although model validation is an important requirement in the development of computational methods, there are cases where it can become problematic. One such case is molecular similarity. Due to its subjective nature, well-defined values of molecular similarity do not exist. Hence, directly assessing the results of similarity calculations is not possible, and indirect methods must be used. These methods are typically based on the SPP noted in Sect. 1.3.1 (see also [13–15]) and assess the recovery rates (or some related measure) obtained from similarity searches of large compound DBs containing known actives [108, 109]. Two such measures are the *recall* and *precision* of compound retrievals given, respectively, by

$$\begin{aligned}\text{Recall} &= \frac{\text{Number of actives retrieved}}{\text{Total number of actives}} = \frac{n_{\text{Act}}^{*}}{n_{\text{Act}}}\\ \text{Precision} &= \frac{\text{Number of actives retrieved}}{\text{Total number of compounds}} = \frac{n_{\text{Act}}^{*}}{n}\end{aligned} \tag{1.44}$$

These measures, although relatively widespread, have a number of deficiencies, one of which is that they do not sufficiently account for “early enrichments” in sets of retrieved compounds. This issue can be dealt with using cumulative recall curves, which plot the fraction of actives against the number of compounds retrieved [108, 109]. These curves are similar to receiver operating characteristic (ROC) curves. Truchon and Bayly [110] have provided a detailed analysis of their application to virtual screening methods.

Significant issues remain that can confound attempts to assess the validity of similarity measures: (1) Untested DB compounds are *assumed* to be inactive, an assumption that is problematic at best. (2) The presence of *activity cliffs* [111–113], which arise when small changes in structure are associated with large changes in biological activity, although rare, represent violations of the SPP giving rise to what Stahura and Bajorath call the “similarity paradox” [114]. (3) The surprising prevalence of *similarity cliffs* [7, 8], which in contrast to activity cliffs occur when small changes in activity are accompanied by large differences in similarity, suggests that active compounds tend to be scattered throughout CSs, although they are likely to be found in multiple clusters of actives, not as singletons, dispersed throughout those spaces.[7] (4) As noted earlier, similarity measures are not invariant to the representation and similarity coefficient used. This lack of invariance leads, either directly or indirectly, to the notion that combining the results obtained from multiple similarity measures, as discussed in Sect. 1.2.3, can yield improved results in molecular similarity analyses.

The prevalence of similarity cliffs noted above also provides a rationale, albeit a tentative one, as to why group fusion (Sect. 1.2.3.2) performs as well does. Numerous analyses by Willett and his colleagues show that it appears to work best with diverse rather than highly similar reference sets [92, 94, 106]. Their conclusion

[7] It should be noted that similarity cliffs are more general than scaffold hops since all scaffold hops do not result in compounds that are highly dissimilar, as may be the case when the scaffolds associated with scaffold hops are approximate bioisosteres or compounds with dissimilar scaffold nonetheless have similar overall structures.

is consistent with the unexpectedly high occurrence of similarity cliffs in pairs of active compounds. In fact, in more than 50 % of the cases where both compounds in a compound pair are active (i.e., $p\mathrm{K}_i \sim 7$), the compounds are also dissimilar [7] (cf. [115][8]).

The significant presence of similarity cliffs suggests that similarity search methods that rely on single active reference compounds, regardless of whether single or multiple similarity measures—as in similarity fusion—are used, will by their very nature miss a significant portion of potentially active compounds because only the top scoring or highest-ranked compounds obtained in similarity searches are typically chosen—compounds located further down the ordered list are routinely ignored.

Group fusion, on the other hand, employs multiple reference actives and, as noted above, performs best when the reference compounds are as diverse as possible. Hence, the dispersion of active compounds is explicitly accounted for by the method, although the available reference set may not, in many cases, provide sufficient coverage of all of the regions of CS that contain active compounds with respect to the given assay, and some actives will undoubtedly be missed.

Because group fusion uses either the MAX rule for similarities or the RRF rule for rankings, compounds located close to the reference compounds are given preference over more distant, less similar compounds, a situation that accords well with the SPP since compounds located close to known actives are more likely to also be active than are less similar compounds. Thus, the performance of group fusion can be rationalized by the significant presence of similarity cliffs in activity landscapes.

1.2.5 Computational Versus Perceptual Aspects Molecular Similarity Measures

The computational methods described above provide algorithms for computing molecular similarities, albeit imperfect ones, due to the inherently subjective nature of similarity. This, however, begs the question as to how these similarity measures accord with the perceptions of chemists, an issue that has been discussed in more detail in several recent publications [10, 34]. An important question in this regard is whether similarity scales used intuitively by chemists agree with those obtained computationally. The answer, as we shall see, is that they do not.

Essentially, all computed similarity values lie on the unit interval [0,1] of the real line (more correctly the unit interval of the rational line). Highly similar molecules have values at the high end of this scale, while dissimilar molecules tend to lie at the lower end. Humans can, in general, assess the similarity of very similar objects, as chemists can assess the molecular similarity of molecules with similar structures. But what happens when molecules become less similar (more dissimilar)? There

[8] Even though the overall percentage of active compounds in large DBs is usually quite small, since most compounds are inactive in a given assay, the fraction of those actives where both compounds of a compound pair are approximately of equal activity can be significant.

is basically no issue with computational similarity measures, but humans, on the other hand, find it increasingly difficult to assess the degree of similarity of highly dissimilar objects. Beyond some point, all that can be said is that the objects are “not very similar,” but the degree of similarity becomes moot. This is also true for chemists’ assessments of highly dissimilar molecules.

Is this difference between computational and perceptual measures of similarity important? Since chemists are unable to perceive low degrees of similarity among molecules, low values of computed similarity do not have any explicit “structural meaning,” at least to chemists. Because of this, it is difficult for chemists to make meaningful structural inferences as would be required when, for example, assessing the diversity of or clustering a compound library, or evaluating compounds for acquisition [116–118]. Computers, on the other hand, are not saddled with this perceptual limitation, and thus can handle similar and dissimilar molecules with equal ease.

Another matter bears on the issue of computation versus perception of structural similarity. In the former case, as described in previous sections, the similarity value obtained depends on the molecular representation used, the weighting of its components, and the similarity coefficient. Changing any or all of these can result in significant changes to the computed similarity values. By contrast, perception of molecular similarity depends on a chemist’s training, experience, and the field of chemistry in which they work. For example, a synthetic organic chemist might focus on likely sites of substitution, a medicinal chemist on the placement and nature of pharmacophoric groups, and a physical chemist on the electron distribution or the energy of a molecule’s highest-occupied and lowest-unoccupied orbitals.

1.3 Chemical Spaces

The amount of chemical information is growing exponentially. Thus, a framework is needed for dealing effectively with the flood of information. The concept of CS provides such a framework. In analogy to mathematical spaces, CSs are specified by a set of molecules and a binary relation that characterizes the relationship of one molecule to another and is typically based on some type of similarity, dissimilarity, or distance measure. Importantly, the notion of CS provides a basis for the well-known SPP that explicitly or implicitly underlies many applications of similarity in chemical informatics (*vide infra*) and is discussed in the following section.

CSs come in three flavors: (1) coordinate based, (2) cell based, and (3) graph or network based. Multidimensional vectors with continuous, real-valued components define the positions of molecules in coordinate-based CSs. The value associated with each of the coordinates is obtained from one of a wide variety of property descriptors discussed in Sect. 1.2.2.1. A simple 3-D example is given in Fig. 1.7a, but since these spaces are generally greater than dimension three, their graphic portrayal requires some type of reduction in the dimensionality of the space. Details of how this can be accomplished will be described in Sect. 1.3.2.

By contrast, compounds in cell-based CSs reside in p-dimensional hypercubes called cells that *partition* the original p-dimensional coordinate-based CS. Cell-based

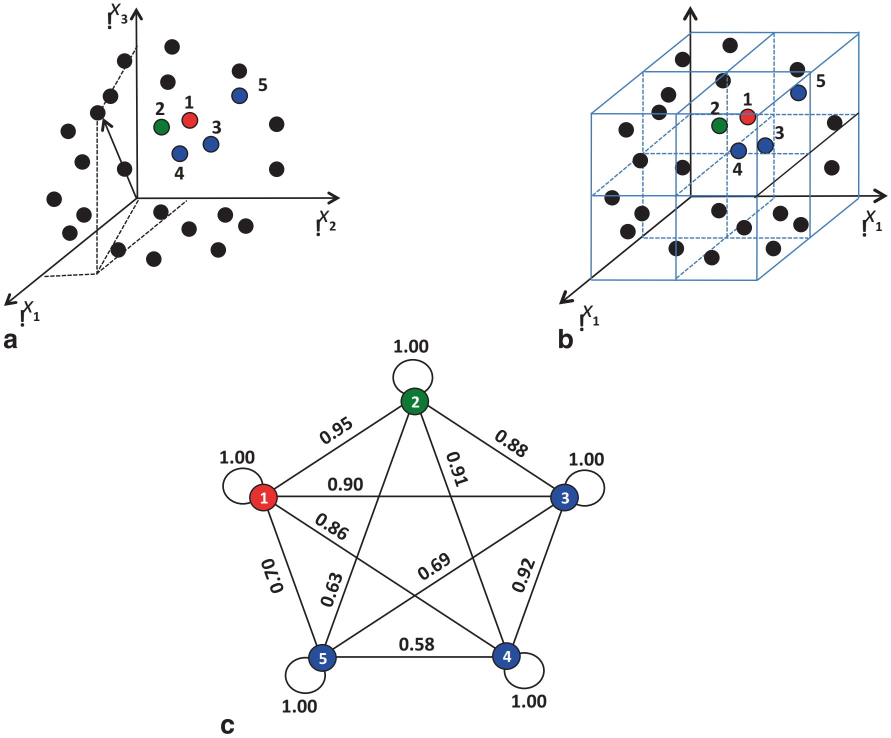


Fig. 1.7 Examples of different CS representations: **a** coordinate-based CS. **b** cell-based CS, and **c** depiction of a complete, self-similar CSN representation of the chemical subspace of the five numbered compounds. The *red filled circle* corresponds to an active compound, the *green filled circle* to its nearest neighbor, and the three *blue filled circles* to next-nearest neighbors

partitioning is a *coarse-grained* approach that lowers the resolution of the space but does not necessarily reduce its dimensionality. Nevertheless, it offers potential advantages for handling a number of procedures commonly carried out in chemical informatics, some of which will be discussed in more detail in Sect. 1.3.3. Figure 1.7b depicts a cell-based CS associated with the coordinate-based space illustrated in Fig. 1.7a. Other types of partitioning methods such as recursive partitioning have also been applied to molecular systems [119–121]. However, it should be noted that recursive partitioning and other tree-based decision methods generally fall in the class of supervised machine learning methods, while cell-based and clustering methods generally, but not always, fall into the class of unsupervised methods.[9]

The third type of CS representation is illustrated by the mathematical graph depicted in Fig. 1.7c called a *reflexive, labeled* or *simple, labeled graph*. The term

[9] Supervised machine learning methods typically try to model the relationship of a set of predictor (independent) variables to a set of known values (e.g., biological activities and/or solubilities) associated with one or more dependent variables. Unsupervised methods only require information associated with predictor (independent) variables (e.g., physicochemical descriptors).

"reflexive" indicates that that each vertex possesses a graph loop, while the term "labeled" indicates that each vertex and edge may be labeled by a set of alphanumeric characters or numbers that describe the properties of these graph entities. In the present application, each vertex corresponds to a molecule and each pair of molecules may or may not be connected by an edge labeled by the value of a pairwise property associated with the two molecules. In contrast to the previous two CS representations, this is a *relational model* that provides a faithful, discrete representation of CSs. More specifically, the edges represent binary relations associated with similarities, dissimilarities, or distances among compound pairs and the nodes are associated with individual compounds. Because CSs typically contain many compounds, the graphs representing them are quite large and generally fall under the rubric "networks." Network research has experienced extremely rapid growth over the past decade in a number of fields from social science [122] to biology and medicine [123–126] as well as in the popular literature [127, 128]. In this regard, the book by Newman not only provides an excellent overview of many aspects of networks but also addresses a number of algorithmic issues associated with them that are critical to their effective application [129].

Although the network model of CS is not in extensive use today, it corresponds closely to the data model of a new graph-based DB technology [130], and thus may provide an additional incentive for adopting this model for future work in chemical informatics. Details of how networks can be applied to the study of CSs are provided in Sect. 1.3.4.

1.3.1 Similarity-Property Principle

The SPP plays a major role in chemical informatics since it provides a crucial link between the similarity of molecules and their corresponding bioactivities or properties. Wilkins and Randic formally described this principle, which now seems intuitively obvious, in a seminal paper published more than three decades ago [13]. Although "similarity" arguments had been advanced in chemistry before this time (cf. [9]), none directly addressed the structural similarity between molecules in a computationally amenable form. In the late 1980s and the early 1990s, the SPP was reiterated [14, 15] and since that time has played a substantive role, explicitly or implicitly, in numerous studies associated with similarity searching and virtual screening.

While the SPP obtains in most cases, there are some notable exceptions such as the presence of activity cliffs [111–113], which arise when pairs of similar compounds exhibit significantly different activities leading to *quasi-discontinuities*[10] in their corresponding CSs [131, 132]. Although statistically rare [7, 8], activity cliffs provide significant SAR information because they afford a means for identifying

[10] Since CSs are inherently discrete, the concept of discontinuity, which applies to continuous systems, is only approximate. Thus, "discontinuities" in these spaces, such as those arising from the presence of activity cliffs, are denoted as quasi-discontinuities.

small structural changes, for example, the presence or absence of a functional group, that are associated with correspondingly large changes in activity.

Another quasi-discontinuous feature occurs in the case of *similarity cliffs* that, in contrast to activity cliffs (*vide supra*), represent compound pairs where small changes in biological activity are associated with large changes in similarity (*vide supra* Sect. 1.2.4). Thus, these cliffs are related to the notion of *target promiscuity* that stands in sharp contrast to the better-known notion of *compound promiscuity* associated with polypharmacologies [124, 133]. The fact that similarity cliffs are the most prevalent feature observed in activity landscapes for active compounds [7, 8] implies that target promiscuity is also more prevalent than heretofore had been assumed. Taken together, both concepts reinforce the idea that compound specificity may be a difficult goal to attain in many instances.

As noted earlier, since similarity measures are not invariant to the representation or similarity coefficient employed, small differences with respect to one measure may not be comparably small with respect to another measure. In such cases, activity cliffs themselves will not be invariant to similarity measure [10, 34, 41], an uneasy state of affairs that raises the question of whether activity cliffs actually exist [134]. Alternative representations based on *matched molecular pairs* (MMPs) have sought to address this question using the 2-D structural representation favored by chemists, but entirely quantitative results have yet to be obtained [135, 136]. Because of its inherent subjectivity, it is unlikely that invariant values (absolute values) of molecular similarity can ever be obtained. Nevertheless, while it may be difficult to quantitate the magnitudes of activity cliffs, there is no doubt that they exist since many examples of "small" structural changes, as perceived by medicinal chemists, have resulted in relatively large activity differences [134].

Based on earlier work by Brown and Martin [17, 18], Martin et al. [137] have provided an updated assessment of the SPP in medicinal chemistry. They examined a large dataset containing the results from more than 100 different HTS assays and concluded that there is only about a 30 % chance that a compound with a Tanimoto similarity value ≥ 0.85 (based on daylight FPs [138]) to a known active is also active, significantly revising an earlier estimation of 80 % [139] (cf. [140]). However, a recent publication [34] has shown that such thresholds may not, in any case, be statistically significant.

Steffen et al. [141] have described a novel approach to the SPP that differs significantly from typical FP methods. In their work, these authors employed a vector representation, where the vector components are categorical variables [41] and are based on the activities of compounds with respect to each one of a fixed set of assays. Hence, the vectors live in "biological activity space" not, as is usually the case, in some form of structure space. This enables the potential identification of compounds with similar biological activity profiles that are structurally dissimilar—compound pairs that fall into this class are related to *similarity cliffs* (*vide supra*) [7, 8]. These authors also showed that representations that included physicochemical or pharmacophoric features were generally better able to retrieve dissimilar compound pairs with similar biological activity profiles. Importantly, this work opens up new possibilities in the study and application of the SPP.

Given the caveats described above, it is important to remember that the SPP is applicable to any type of CS regardless of how it is represented (*vide infra*). Today, the SPP is applied explicitly in many areas of chemistry, but particularly in medicinal chemistry. It might be said that the SPP, whether it is used explicitly or implicitly, is one of the foundations of medicinal chemistry.

1.3.2 Coordinate-Based CSs

The most common representation of coordinate-based CSs is as a set of points, each representing an individual molecule, embedded in a multidimensional Euclidean space much like the stars and planets in our galaxy. In general, *p*-dimensional Euclidean spaces have *p* orthogonal coordinate axes, and each of the *n* points occupying the space is described by a *p*-dimensional vector as that given in Eq. (1.27). The set of row vectors can then be combined into the $n \times p -$dimensional data matrix given in Eq. (1.28), which contains the molecular and/or chemical information associated with the entire set of compounds. The relationship between any compound pair can be assessed in several ways: (1) by any of the vector-based similarity coefficients described in Eqs. (1.34)–(1.36**)**, (2) by any of the corresponding dissimilarity coefficients described in Eqs. (1.37) and (1.39), or (3) by the Euclidean distance in CS between two molecular feature vectors as described in Table 1.2 and Eq. (1.38).

Figure 1.7a provides an illustration of a simple model 3-D CS. The five color-coded compounds are, respectively: Cpd-1, an active colored in red; Cpd-2, its nearest neighbor colored in green; and Cpd-3, Cpd-4, and Cpd-5, the three next nearest neighbors, colored in blue, are ordered with respect to decreasing similarity (or increasing dissimilarity or distance) with respect to Cpd-1. Thus, Cpd-3 is nearer to Cpd-1 than Cpd-4, which is closer than Cpd-5.

Figure 1.8a portrays a 3-D projection of a real, six-dimensional (6-D) 3-D BCUT CS; additional details on its construction are supplied in Sect. 1.3.3.2. The projection is with respect to the three most significant BCUT descriptors that are derived from the electronic (“Elec”), hydrophobic (“HPhob”), and hydrogen-bonding (“HBond”) features of atoms (see Sect. 1.2.2.1 for a more detailed description of BCUT descriptors). A diverse set (“Diverse”) containing approximately 175,000 compounds is depicted in yellow; a combinatorially generated set (“Combi”) containing approximately 150,000 compounds constructed from a set of 40 different scaffolds, is depicted in red. It is clear from the figure that Combi, which is of nearly comparable size to Diverse, covers only a small fraction of the CS covered by the latter. Figure 1.8b shows a magnified version of the CS shared by both collections.

The fact that many data spaces including CSs possess more than three dimensions has, over the years, generated a significant amount of effort in the development of dimensionality reduction techniques. There are three main reasons for reducing dimensionality. The first and most obvious is that graphical depiction of the space is restricted to three or fewer dimensions. The second and more important reason is due to the “curse of dimensionality” [142] that occurs because the data

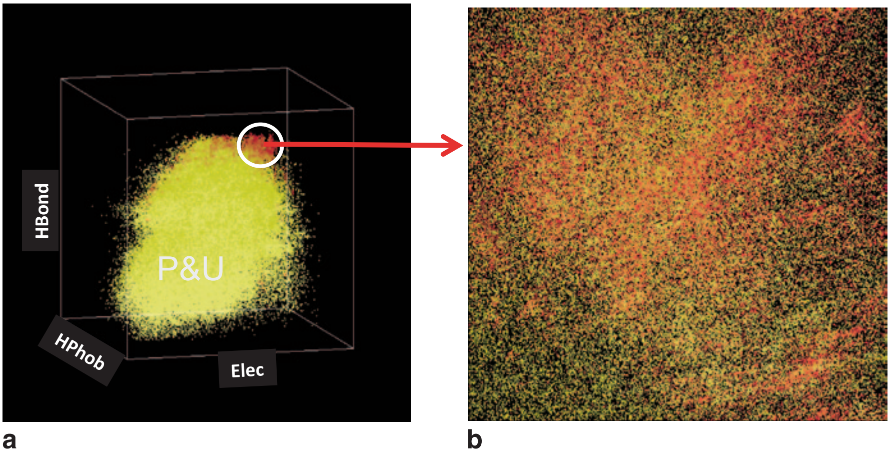


Fig. 1.8 **a** Example of a three-dimensional projection of a six-dimensional 3-D BCUT CS containing *ca.* 175,000 molecules depicted in *yellow*, and a combinatorial library of *ca.* 150,000 molecules depicted in *red*. **b** Magnified version of the region of CS shared by both sets of molecules (See Section *2.2.1* for description of BCUT descriptors). (Figure kindly provided by Veer Shanmugasundaram)

distribution becomes more sparse as the dimension of the space increases. Thus, in order to ensure balanced or comparable coverage of the resulting higher-dimensional space requires an increase in the amount of data, which becomes more difficult to achieve as the dimension increases. Higher-dimensional spaces can also exhibit idiosyncratic behaviors that are difficult to comprehend [143]. The third reason is that the intrinsic dimension of the data may be considerably lower than its apparent dimension and may in some cases be confined to a non-Euclidean subspace, which could also be nonlinear. As discussed in Sect. 1.3.2.3, distances between points in non-Euclidean subspaces are generally different than they are in the Euclidean space in which they are embedded.

1.3.2.1 Coordinate-Based CSs Derived from Structural FPs

Constructing coordinate-based CSs from low-dimensional vector representations, which is relatively straightforward, is exemplified by BCUT descriptors described in Sect. 1.2.2.1. Figure 1.8 depicts an example of a 3-D BCUT chemical subspace projected from the original 6-D BCUT CS. Today, a common means for representing molecules is by their structural FPs. However, their direct use in the construction of coordinate-based CSs is beset by a number of problems that include: (1) they are generally of very high dimension, usually in the range of ~ 150–2000, and hence are plagued by the curse of dimensionality [142] and (2) their coordinates are generally binary or integer valued and thus are not compatible with the types of continuous, real-valued CS representations described above. Nevertheless, structural FPs can be transformed into continuous, real-space coordinates in a number of ways usually through the computation of some pairwise measure that characterizes the relationships among the molecules of the set. These relationships are typically associated with the similarity or dissimilarity coefficients described in Sect. 1.2 or with some type of CS distance such as the Hamming distance [60].

A distinct advantage of this approach is that any type of representation can be used that affords a means for computing a similarity, dissimilarity, or distance measure. For example, chemical graphs [16], which cannot be treated using a purely coordinate-based approach, can be handled in a straightforward, albeit somewhat computationally demanding, manner [41]. Recent work on graph-based Kernel methods provides a novel means for extending and generalizing methods for computing similarity coefficients [144].

Given that a matrix of similarity, dissimilarity, or distance values can be computed for each unique pair of molecules, the question now becomes, "How can this array of values be transformed into a set of coordinates that define the positions of molecules in a coordinate-based CS?" In this regard, most efforts in chemical informatics have generally focused on five main techniques: (1) principal component analysis (PCA) [145], (2) principal coordinate analysis (PCoA) [145], (3) multidimensional scaling (MDS) [146], (4) nonlinear mapping (NLM) [147], and (5) factor analysis [148]. All five methods provide the means for constructing low-dimensional representations of CSs. A recent review by Shanmugasundaram and Maggiora [41] provides additional details and references to these methods.

Although any of the five methods would suffice, PCA, a method used in many chemical informatics applications, will be employed here as an example of how CSs can be constructed from several varieties of structural FPs. Consider the similarity coefficient values of a set of n molecules computed with respect to some type of structural FP that generates an $n \times n -$ dimensional symmetric matrix of similarity coefficients[11]

$$\mathbf{S}_{n\times n} = \begin{bmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,j} & \cdots & s_{1,n} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,j} & \cdots & s_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i,1} & s_{i,2} & \cdots & s_{i,j} & \cdots & s_{i,n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,j} & \cdots & s_{n,n} \end{bmatrix} \tag{1.45}$$

where $s_{i,j}$ corresponds to any of the similarity coefficients described in Sect. 1.2. There is no need to scale these values since they are all on the same scale and lie on the unit interval [0,1] of the real line.

Although the matrix does not have the form of a typical data matrix, the similarity values can, nevertheless, be thought of as descriptor values. Consider, for

[11] In mathematics these are generally called Gram matrices and in statistics are usually called association matrices.

example, the *i, j*th element of $\mathbf{S}_{n\times n}$, which can be interpreted as the similarity of the *i*th molecule in the set of *n* molecules with respect to the *j*th "descriptor molecule". In this case, the *n* "descriptor molecules" are taken from the same set of *n* molecules under study—a generalization of this approach was recently described [149]. As was suggested by Kruscal [150], square symmetric matrices such as $\mathbf{S}_{n\times n}$ can be handled in exactly the same manner that general data matrices are treated using PCA:

$$\mathbf{S}_{n\times n} \Rightarrow \underset{\text{Mean center}}{\bar{\mathbf{S}}_{n\times n}} \Rightarrow \underset{\text{Compute covariance matrix}}{\mathbf{C} = \tfrac{1}{n-1}\bar{\mathbf{S}}^{\mathrm{T}}_{n\times n}\bar{\mathbf{S}}_{n\times n}} \Rightarrow \underset{\text{Diagonalize } \mathbf{C}}{\mathbf{V}^{\mathrm{T}}\mathbf{C}\mathbf{V} = \Lambda} \tag{1.46}$$

the eigenvalues

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \tag{1.47}$$

which are ordered from largest to smallest, are related to the variances in the new, transformed coordinate system[12]

$$\mathbf{Z}_{n\times n} = \bar{\mathbf{S}}_{n\times n}\mathbf{V}_{n\times n} \tag{1.48}$$

such that the percent of the total variance corresponding to the *i*th eigenvalue is given by

$$\text{Percent-Variance}(\lambda_i) = \frac{\lambda_i}{\sum_{j=1}^{n}\lambda_j}\times 100 \tag{1.49}$$

Thus, to graphically depict a CS in three dimensions, the transformed coordinates associated with the first three eigenvalues will suffice, i.e.,

$$\mathbf{Z}_{n\times 3} = \bar{\mathbf{S}}_{n\times n}\mathbf{V}_{n\times 3} \tag{1.50}$$

Note, however, that the entire mean-centered similarity matrix $\bar{\mathbf{S}}_{n\times n}$ is required.

Although this procedure provides a reasonably straightforward approach to the construction of low-dimensional CSs, the number of compounds that can be handled is somewhat limited because determining the transformed coordinates requires diagonalization of the $n\times n$ covariance matrix, which becomes difficult for $n > 2500$, although there are ways that this limitation can be overcome, for example, by using real time PCA [151].

Figure 1.9 shows examples of CSs constructed with respect to four different binary FP representations using the similarity-based PCA procedure described in the previous section. The first two examples are based on atom pair and MACCS key FPs that were discussed in some detail in Sect. 1.2.1.1. Of the latter two, both

[12] Note that the coefficient $(n-1)^{-1}$ would, if ignored, merely scale the eigenvalues by $n-1$; the eigenvectors are unaffected.

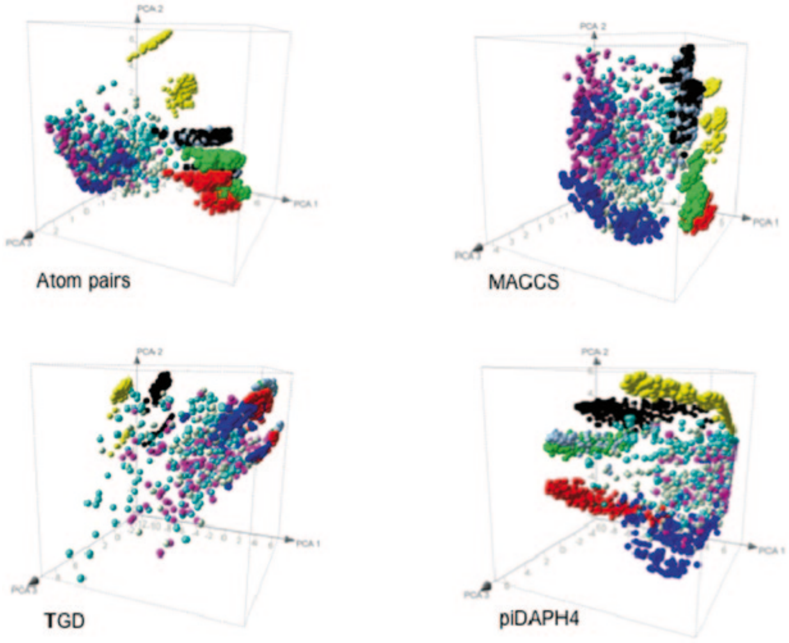


Fig. 1.9 Depictions of CSs generated from Tanimoto similarity coefficients computed with respect to binary FPs associated with four different types of descriptors—APF, MACCS key, TGD, and piDAPH4. (Adapted from Medina-Franco & Maggiora, Molecular Similarity Analysis [10])

of which are available in molecular operating environment (MOE) [152], TGD FPs are similar to those in atom pairs, while the piDAPH4 are related to FPs whose components are 3-D pharmacophores [153]. Hence, in contrast to the first three, piDAPH4 FPs contain some 3-D structural and stereochemical information. The Tanimoto similarity coefficient given in Eq. (1.8) was used to compute the similarity value in all four cases.

A total of 2250 molecules comprising nine classes of 250 molecules each were considered. The molecules in each class are color coded as follows: approved drugs (cyan), natural products (light green), a general screening collection from two vendors (magenta), compounds targeted to adenosine receptors (blue), and five in-house combinatorial libraries from the Torrey Pines Institute for Molecular Studies (depicted red, yellow, green, black, and light blue). The first three PCs account for 80.8, 85.9, 90.3, and 73.0 % of the total variance in the data associated with the atom pair, MACCS key, TGD, and piDAPH4 FPs, respectively.

Although some of the variance in the data is not accounted for in the 3-D plots, a significant portion of it is. Hence, it is possible to draw some conclusions, albeit qualitative ones, from the distribution of compounds associated with the four different FP representations. It is quite obvious from the figure that the four different FP representations lead to dramatically different graphical portrayals of the CS distributions of the same set of compounds, a not unexpected but visually dramatic example of the non-invariance of similarity measures and its consequences. Interestingly, in some cases, substantial differences arise even within individual compound classes, as shown, for example, by the class of approved drugs colored in cyan. Of even greater interest is the graphical depiction in Fig. 1.10 of the distribution of the same set of compounds with respect to similarity fusion based on the mean of the similarity values (see Table 1.3). The results depicted in Fig. 1.10 differ significantly from any of those depicted in Fig. 1.9, which are based on the values of individual, "unfused" similarity measures obtained with respect to four different binary FPs.

It is important to note that graphical depictions described in this section are meant primarily as a means for enhancing intuition about the relationships among molecules in CSs. If quantitative analyses are required, detailed computations can be carried out using the full, multidimensional representation of the molecules in a dataset, as noted earlier.

1.3.2.2 Non-Euclidean Coordinate-Based CSs

The fact that CSs must have fewer than four dimensions for their graphical depiction is obvious. A less-well-known and much more subtle point is that high-dimensional data, in general, and CSs, in particular, may lie on lower-dimensional curved (i.e.,

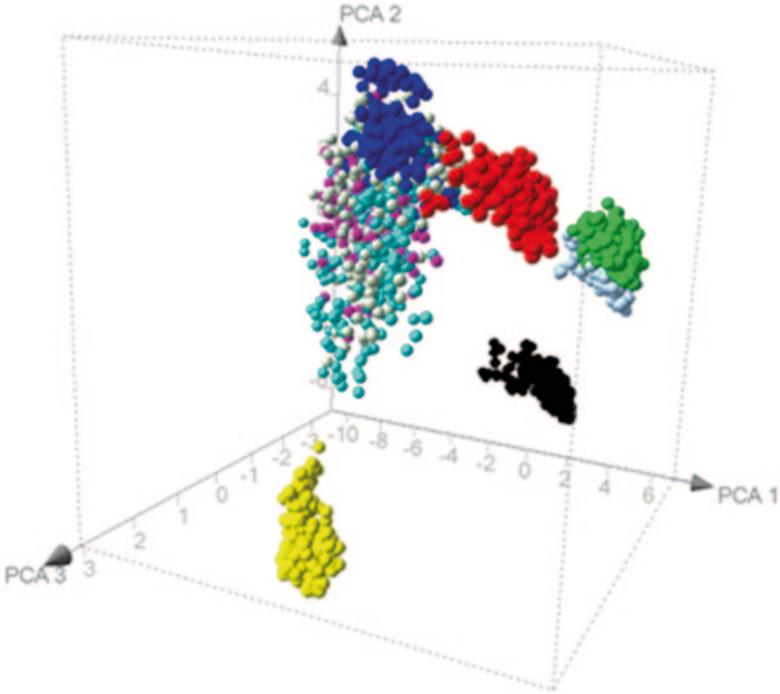


Fig. 1.10 Depiction of a CS generated from mean fusion of similarity values obtained from Tanimoto similarity coefficients computed with respect to binary FPs associated with APF, MACCS key, TGD, piDAPH4 descriptors. (Adapted from Medina-Franco & Maggiora, Molecular Similarity Analysis [10])

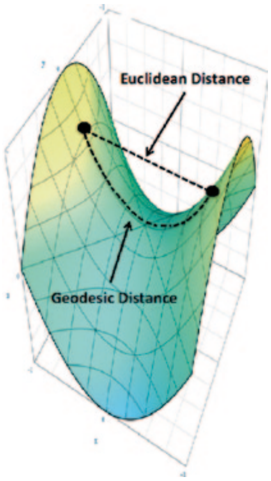


Fig. 1.11 Example of a Euclidean distance and the corresponding geodesic distance of two compounds in a model CS. The surface rendition is by Sam Derbyshire, http://creativecommons.org/licenses/by-sa/3.0

non-Euclidean) manifolds that are embedded in higher-dimension Euclidean spaces (*vide supra*). A simple example is given in Fig. 1.11, which depicts a 2-D hyperbolic manifold embedded in a 3-D Euclidean space. The important point is that the distance between points A and B depends on the space in which the distance is being evaluated. In the example, the Euclidean ("straight line") distance is clearly less than the geodesic distance measured along curved surface of the 2-D manifold. Thus, molecules A and B are judged more similar if considered in Euclidean CS than if their similarity was assessed on the 2-D manifold defined by the hyperbolic surface depicted in Fig. 1.11 that more accurately represents the data (in this toy example).

Figure 1.12 provides a more "down to Earth" example that clearly illustrates the difference between the two distance measures. In this case, the Euclidean distance is given approximately by the air miles between the American cities of Seattle, Washington and Miami, Florida which is about 2730 miles. By contrast, the geodesic distance between these two cities, measured along the US highway system is about 3300 miles, which represents about a 20 % increase in miles by car.

The paper by Agrafiotis and Xu provides a number of examples illustrating geodesic distances [154]. Although these authors published two more papers on this subject [155, 156] very little else has been published in the chemical information literature. This is obviously an important area of future research since it is one of several factors that can significantly influence the computed values of CS distances and, hence, the inferences that can be made about the compounds in a CS.

Lastly, it is well to point out that similarity values that lie on the unit interval [0,1] of the real line can be obtained by transforming Euclidean distances, d, or non-Euclidean geodesic distances, $\hat{d}$, using any one of a number of different mathematical expressions, one possibility being

$$s_{i,j} = 1/[1+\eta \cdot \widehat{d_{i,j}}], \tag{1.51}$$

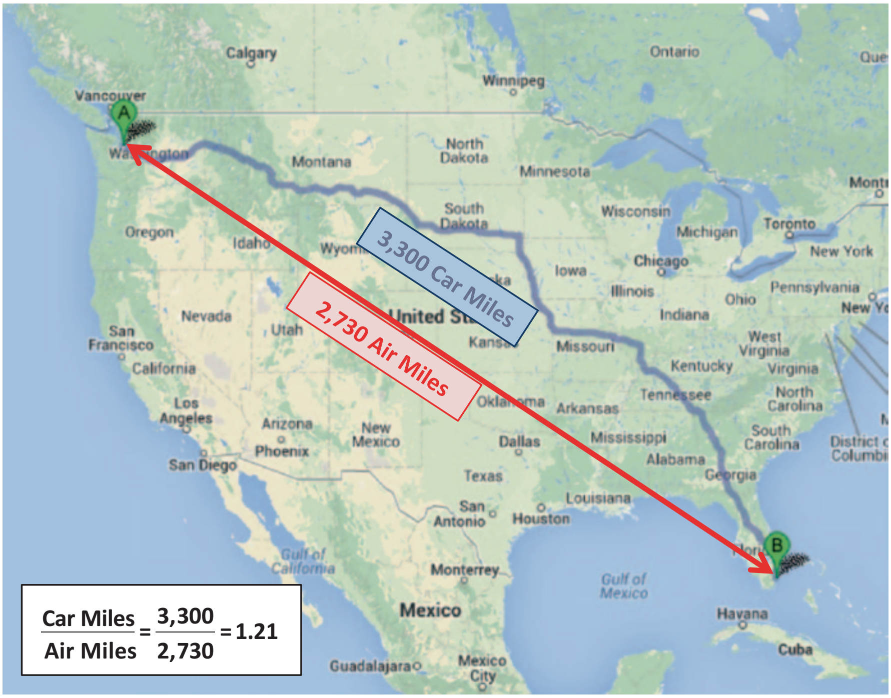


Fig. 1.12 Car (*blue-grey*) vs. air (*red*) routes from Seattle, Washington to Miami, Florida. (Adapted from Google Maps)

where the parameter $\eta > 0$ controls the rate at which the similarity value changes as a function of distance.

1.3.3 Cell-Based CSs

Cell-based partitionings of CSs [76, 157] are identical to partitions of mathematical spaces into families of nonintersecting subsets that cover the spaces. Thus, the set of N_{cells} cells that constitutes a cell-based CS is given by:

$$\mathrm{C}_{\text{cells}} = \left\{ \mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_i, \ldots, \mathrm{C}_{N_{\text{cells}}} \right\} \tag{1.52}$$

and satisfies

$$\begin{aligned} &\mathrm{C}_i \cap \mathrm{C}_j = \varnothing \text{ },1 \le i, j(< i) \le N_{\text{cells}} \quad \text{(Non-Intersecting cells)} \\ &\bigcup_{i=1}^{N_{\text{cells}}} \mathrm{C}_i = \mathrm{C}_{\text{cells}} \text{ (Set Cover)} \end{aligned} \tag{1.53}$$

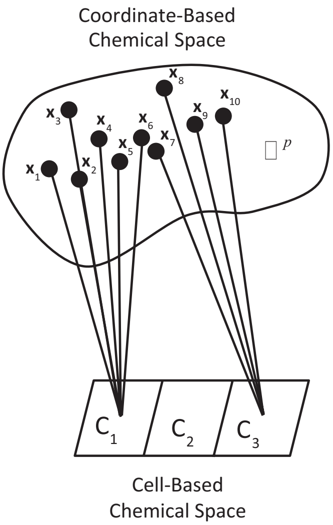


Fig. 1.13 Schematic depiction of a many-to-one set-valued mapping. Note that most cells are not generally occupied in cell-based CSs (see text for additional details)

Each cell corresponds to an *equivalence class*, and the molecules within it are hence, in some fashion at least, equivalent. The *many-to-one set-valued mapping*[13] depicted in Fig. 1.13 takes molecules in a *p*-dimension coordinate-based CS to one of the cells of the corresponding cell-based space, i.e.,

$$\Phi : \mathbb{R}^{p} \rightarrow \mathrm{C}_{\text{cells}} \tag{1.54}$$

Thus, the location of compounds in cell-based CSs is given in two ways, namely, by their coordinates in the underlying coordinate-based CS, and by the address of the cell in which they reside. Figure 1.13 also shows that some cells in cell-based spaces are empty since only 15–20 % of the cells in cell-based CSs are typically occupied. It is also interesting to note that cell-based CSs are very similar to the multi-way contingency tables used in many statistical applications [158], except for the fact that contingency tables rarely have cells with zero values.[14]

The procedure for constructing virtually all cell-based CSs is basically a two-step process:

- Generation of an appropriate low-dimensional coordinate-based CS
- Binning each of the axes of that space in such a way that the occupancy of the bins optimally covers the CS

The first and perhaps most important step in the process is the selection of suitable sets of reference compounds and descriptors, since they both play major roles in

[13] In function notation, the mapping in Eq. (1.54) is given by $\Phi(\mathbf{x}_i) = \mathrm{C}_k,\ i = 1,2,\ldots,n;\ k = 1,2,\ldots,N_{\text{cells}}$.

[14] Note that there are a number of "correction factors," such as the well-known Laplace correction, that can be applied to the cells of a contingency table to correct for empty cells.

determining the nature of the CSs ultimately generated. While it is well appreciated that descriptor selection is important, the role played by the reference set of compounds is perhaps less well appreciated but is nonetheless crucial to the final form of the CS generated. Potential compound sets include corporate compound collections, publically available collections [25] such as ChEMBL [19], PubChem [20], ChemDB [21], and DrugBank [22], or sets of compounds suited to some specific tasks. In the latter case, for example, if the goal is to compare two large sets of compounds, it is desirable to combine the sets since the resulting CS will be more "balanced" and, hence, will take better account of the influence that molecular features missing in one of the two collections may have on the overall representation of the resulting CS. Alternatively, if the goal is to generate diverse subsets for an HTS campaign, the corporate compound collection from which the sample will be drawn, may be the best choice. These are just two of the many possibilities that can be considered, some of which will be presented in the sequel.

The second step in the process involves binning each axis of the coordinate-based CS yielding a total number of cells given by

$$N_{\text{cells}} = N_{\text{bins}_1} \times N_{\text{bins}_2} \times \cdots \times N_{\text{bins}_p} \tag{1.55}$$

As an example, consider a typical 6-D coordinate-based CS with seven bins per axes, which will generate a cell-based CS containing 117,649 cells. Although bins generally are of equal size on each axis, this is not required as discussed by Bayley and Willett [159]. Choosing an appropriate number of bins per axis is also important: If the number is too large, numerous cells will be unoccupied—normally a number of "occupied" cells around 15–20% appears to be reasonable. In this regard, it is important to note that in many types of cell-based analyses, including the above, the specific number of compounds in a given cell is not enumerated, only if the cell is occupied by at least some number of compounds (usually one) called the *cell occupancy threshold value*.[15]

Lastly, while cell-based CSs used in cheminformatic studies are generally partitioned into hypercubes, other possibilities exist that may offer more effective ways to partition these spaces. Rush [160] has mathematically explored some of the possibilities, but practical applications in chemical informatics have not to my knowledge been carried out to date.

Figure 1.7b portrays a model cell-based CS for the same set of compounds depicted in Fig. 1.7a. Although this example is oversimplified, cell-based CSs, nevertheless, are typically around 3-D to 6-D. Cpd-1, the active compound indicated by the red dot, its nearest-neighbor Cpd-2 indicated by the green dot, and two of its next nearest neighbors, Cpd-4, and Cpd-5 indicated by the blue dots, all reside within the same cell. Hence, from a cell-based perspective, all four compounds are considered to be roughly equivalent. On the other hand, Cpd-3, which is nearer to Cpd-1 than either Cpd-4 or Cpd-5, resides in a neighboring cell, and thus, from a

[15] A similar situation exists in the case of threshold graphs obtained from labeled graphs when the edge values exceed some threshold value. Details of this are described in Sect. 1.3.7 on graph-based CSs.

cell-based perspective, is not considered to be equivalent to any of the compounds in the neighboring cell. This illustrates one of the limitations of the cell-based approach, which does not *explicitly* employ the concept of nearest neighbor cells, although the position of compounds in the underlying coordinate-based CS does afford the possibility for identifying nearest neighbors.

Clustering provides an additional way to partition CSs into a set of nonintersecting subsets that cover the space [161]. Although clustering methods have some advantages over cell-based partitioning, they are difficult to apply to datasets as large as those that can be handled relatively easily using a cell-based approach. For example, the addition of large numbers of new molecules can significantly alter clusterings. This is not a problem in the cell-based case since the CS partitioning scheme is effectively compound independent—adding new compounds does not change the partitioning scheme. Moreover, many methods such as k-means clustering require specification of the number of clusters and hierarchical methods produce similarity (or distance)-dependent clusterings [161]. Lastly, because the clustering methods are a vast subject, even when only considered with respect to cheminformatics applications, no further discussion on this topic is provided in this work.

1.3.3.1 Representations of Cell-Based CSs

The BCUT descriptors described in Sect. 1.2.2.1 have proved to be a popular choice for directly constructing low-dimensional CSs. There are, of course, many other types of suitable descriptors that, in many cases, cannot be used directly since they lead to spaces whose dimension are too high. This can be ameliorated, as discussed by Xue, Stahura, and Bajorath [157], using a dimensionality reduction technique such as PCA.

The power of the cell-based description lies in its ability to simplify the representation of CS, and thus to enhance the speed at which a number of the tasks, such as compound acquisition [162], diversity analysis [163], comparison of compound collections [77], and LBVS [164] can be performed. But the enhanced speed comes at a cost, which may or may not, significantly impact the results obtained. As discussed above, the cell-based partitioning leads to a coarse-grained representation of CS and, importantly, can introduce significant effects at cell boundaries. For example, molecules located near a common boundary in adjacent cells are generally more similar to each other than to many other molecules in their own cells (cf. Figs. 1.7 and 1.13). Obviously, this can lead to significant bias depending on the actual (not cell based) distribution of compounds in the CS, a problem that is also encountered in a number of clustering methods.

1.3.3.2 Example of Cell-Based CSs

The CS was constructed by combining the four compound collections given in Table 1.4 into a single, large collection. Determining the optimal set of 3-D BCUT descriptors for that augmented collection yielded a 6-D CS upon which all subsequent analysis is based. Each axis was then partitioned into seven bins, giving a total of 117,649 cells in the 6-D space.

The difference between the Diverse and Combi collections depicted graphically in Fig. 1.8 is verified. Several key features in the table supporting this conclusion are the comparative number of occupied cells (18,731 and 2434, respectively) and the average cell occupancies (9.4 and 61.5, respectively), all of which clearly point to the more restricted and dense distribution of compounds in Combi compared to that in Diverse. The MDDR collection exhibits similar behavior to that of Diverse, although the absolute values of the cell-based parameters are somewhat lower than those of Diverse, which is not surprising given that Diverse is nearly twice as large as MDDR. Micros is a small, diverse collection of known drugs and related substances. Given its size, it nonetheless is relatively diverse since only slightly more than one compound on an average occupies each of the 516 occupied cells. On the other hand, its 516 cells occupied cells are almost insignificant when compared to the 18,371 occupied cells in Diverse. Moreover, each occupied cell in Micros contains on an average only 1.3 compounds, which again pales in comparison to Diverse's average cell occupancy of 15.6.

These data illustrate two important points about diversity. First, small compound collections, which may be relatively diverse with respect to their own set of compounds, may not in an absolute sense contain anywhere near the diversity that can *potentially* be obtained from much larger compound collections. Second, while diversity may confer some advantage in identifying active compounds in HTS campaigns, if the diversity is sparsely distributed the chance of identifying actives is significantly diminished even if the diversity is widespread in a large compound collection. This follows from the fact that in a given assay the percentage of actives within "active regions" of CS is still surprisingly small, generally around 10–15 % or less.

The cell-based CS data summarized in Table 1.4, while helpful, are not sufficiently detailed to address more specific questions regarding the similarity or difference between different compound collections. This is remedied in Sect. 1.3.5.1 where details for comparing compound collections are described.

Table 1.4 Summary of compound collections in six-dimensional 3-D BCUT chemical space with seven bins per axis (total cell count = 117,649)

Compound collection	Number of compounds	Number of occupied cells	Percent occupied cells	Average cell occupancy	Largest cell population
Diverse[a]	173,375	18,371	15.6	9.4	738
Combi[b]	154,474	2434	2.1	61.5	5694
MDDR[c]	97,409	10,203	8.7	8.5	349
Micros[d]	799	516	0.4	1.3	7

[a] Subset of diverse compound collection (see text)

[b] Combinatorial chemistry library (see text)

[c] Subset of MDDR collection—Molecular Drug Data Report (MDDR), Version 2005.2; Symyx Software: San Ramon, CA, 2005

[d] Small discovery oriented library—MicroSource Discovery Systems, Inc., Gaylordsville, CT 06755

1.3.4 Chemical Space Networks

In addition to the coordinate and cell-based representations just described, CSs can also be represented by *mathematical graphs*. Such graphs provide information that is comparable to that provided by similarity, dissimilarity, or distance matrices and, as will be seen in the sequel, afford an intuitive as well as solid conceptual basis for analyzing many relationships among the compounds populating CSs. Since compound collections can be quite large, their corresponding graphs are also quite large and generally fall under the rubric of "Networks." The development and application of network theory, which has burgeoned over the two decades, has been applied in numerous fields, including social science, physics, computing, biology, and medicine. A number of "chemically oriented" examples have been reported (see e.g., [123–126, 165–167]), and five papers describing the application of networks to the analysis of compound collections have been published [168–172]. An investigation that examines power laws in chemical systems, as do several of the just cited publications, has also been published. However, it does not directly address issues related to similarity-based networks that describe compound collections [173].

The present section provides a number of examples that elucidate the underlying features of networks such as their patterns of vertex connectivity. An understanding of these feature patterns is required in order to comprehend the nature of the large, complex networks such as those needed to represent CSs; because these networks are large, their feature patterns are usually analyzed in statistical terms. An important aspect of the network representation of CSs is that it facilitates navigation of those spaces since there are powerful graph-based network algorithms for determining paths between vertices [129] in contrast to the situation in more traditionally represented CSs [174].

In order to facilitate understanding of networks, a number of simple examples based on the graphs depicted in Figs. 1.7c and 1.14 are presented in the following sections. These examples, though simple, illustrate a number of the most important network features needed to interpret the statistical data and to understand the nature of the CSs being analyzed.

1.3.4.1 Simple Example of a CS Network

As an illustration of the basic features of graphs, consider the *reflexive, labeled graph* $\hat{\mathcal{G}}$ depicted in Fig. 1.7c that represents the similarity relations among "hypothetical" compounds 1–5 depicted in Fig. 1.7a, b. A compound identifier, which is a number in the present case, labels each vertex and a similarity value labels each edge of $\hat{\mathcal{G}}$. Since the vertices represent distinct molecules they are distinguishable, a feature that influences the statistical mechanical features of networks (*vide infra*) [175]. As noted earlier, the graph is reflexive because each vertex has an associated graph loop labeled by the value of the self-similarity[16] of the molecule that

[16] Self-similarity is the similarity of the molecule with itself, and thus, its value is always unity. Graphs without self-loops and multiple edges between vertices are also called simple graphs.

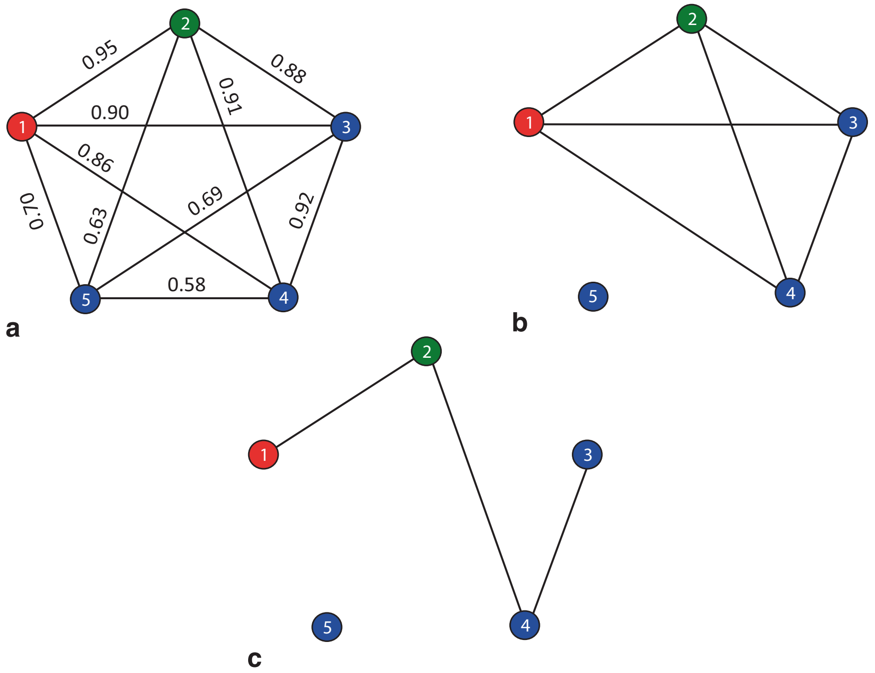


Fig. 1.14 Other CSNs related to that depicted in Fig. 1.7c: **a** simple, complete CSN, **b** threshold CSN ($S_t > 0.85$); the CSN linking compounds 1–4 is a complete *subgraph/network* called a *clique*, and **c** threshold CSN ($S_t > 0.90$); while compounds 1–4 are still linked they no longer form a clique

corresponds to that vertex. In most practical implementations, edges corresponding to self-similarities are omitted for clarity (*vide infra*). Since similarity coefficients are generally symmetric, i.e., $S(i, j) = S(j, i)$, the edges of the corresponding graph do not have directionality. Hence, the networks typically employed can be classified as *undirected, unlabeled*, and *simple networks*.

There are, however, cases when the use of directed graphs may be desirable as in the representation of activity cliffs [112] or where asymmetric similarity coefficients such as those given in Eqs. (1.6) and (1.11)–(1.13) are employed. Graphs where each vertex is connected to every other vertex connected are called *complete*. Thus, a complete graph with n vertices has $n(n-1)/2$ edges, and each vertex has $n-1$ edges called its vertex degree.

The similarity matrix given in Eq. (1.56) contains the same information as $\hat{\mathcal{G}}$ in Fig. 1.7c:

$$\mathbf{S} = \begin{bmatrix} 1.00 & 0.95 & 0.90 & 0.86 & 0.70 \\ 0.95 & 1.00 & 0.88 & 0.91 & 0.63 \\ 0.90 & 0.88 & 1.00 & 0.92 & 0.69 \\ 0.86 & 0.91 & 0.92 & 1.00 & 0.58 \\ 0.70 & 0.63 & 0.69 & 0.58 & 1.00 \end{bmatrix} \tag{1.56}$$

Hence, the similarity matrix provides a means for treating graphs algebraically [176]. For example, the eigenvalues associated with the matrix representations characterize a variety of graph invariants that have seen many useful applications in chemical graph theory [16], and although they have not yet been applied extensively in the study of CSs, they, nonetheless, have the potential to provide new and interesting insights in graph-based CSs.

The example in Fig. 1.7c is, of course, a great simplification of "real" CSs that may contain millions of vertices each corresponding to a specific molecule and billions of edges linking the pairs of vertices each labeled by an appropriate similarity, dissimilarity, or distance value. In this work, the networks are called "CS networks" (CSNs) to emphasize their relationship to CSs. Hence, the graph in Fig. 1.7c can be described as a *complete-reflexive-labeled* CSN. The reflexive character of the graph is captured by the values of diagonal elements of similarity matrix, $S(i,i)=1, i=1,2,\ldots,n$. Since the self-similarities do not add any new information since they are all the same and of value 1.00, graph loops are routinely omitted yielding the simple graph $\hat{\mathcal{G}}$, as illustrated in Fig. 1.14a. Such networks will be called *complete* CSNs since each vertex is connected to every other vertex except itself as the graph loops have been removed.

Because CSs are so large, their graphical display as CSNs can become visually "noisy" and difficult to comprehend for all but the smallest sets of compounds. Nevertheless, as in the case of the coordinate-based portrayal of CSs, the graphical depictions are only meant to provide an intuitive feel for the underlying relationships associated with the CSN of a large compound collection. Alternative ways exist, however, for characterizing and handling the information contained in CSNs. Because matrices can provide faithful representations of graphs and networks, this affords the possibility that many powerful algebraic techniques can be applied to their analysis [177]. Algorithmic techniques, some but not all of which are based on the properties of graph matrices, have provided numerous other ways for analyzing the properties of graphs and networks. However, because of their size and complexity, information on the characteristic features of networks obtained using these methods is commonly reported in terms of the statistical properties of the features, as will be described in Sect. 1.3.5.1 [129, 178].

All of the existing publications that describe applications of networks to CS analysis [168–172] do not use labeled graphs or networks, but rather rely on simpler entities called *threshold graphs*, which are generated by keeping only those labeled edges whose values satisfy some threshold as illustrated in Fig. 1.14b, c. In the first case, shown in Fig. 1.14b, a similarity threshold value of $S_t > 0.85$ is used. Vertex 5 is now isolated from the vertices 1–4, which remain fully connected, and thus form a complete subgraph of the original graph called a *clique*. Figure 1.14c provides another example based on a higher threshold value of $S_t > 0.90$. Not surprisingly, fewer edges remain, and although vertices 1–4 are still connected, they no longer form a clique.

An important type of matrix that plays a role in many procedures designed to determine graph/network properties is the *adjacency matrix* of mathematical graphs and networks. The adjacency matrix corresponding to the CSN in Fig. 1.14b is given by

$$\mathbf{A}_{0.85} = \begin{pmatrix} \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} & 0 \\ 0 & [0] \end{pmatrix}, \tag{1.57}$$

Where

$$a_{i,j} = \begin{cases} 1 \text{ if an edge exists between Cpd-}i\text{ and Cpd-}j \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases} \tag{1.58}$$

As noted above, the subset of compounds $\{\text{Cpd-1,Cpd-2,Cpd-3,Cpd-4}\}$ forms a complete subgraph of the threshold graph called a *clique,* i.e., $\mathcal{H}_{0.85} \subset \mathcal{G}_{0.85}$. Thus, the four compounds are all linked in the threshold CSN, while Cpd-5 is an isolated vertex as reflected by the block diagonal structure of the adjacency matrix in Eq. (1.57). Because of the block diagonal structure, each block can be treated independently of the others, a form of dimensionality reduction.

If the threshold is raised, to say $S_t > 0.90$, the subset of compounds remains linked, but the subgraph induced by the higher threshold $\mathcal{H}_{0.90}$ no longer forms a clique and $\mathcal{H}_{0.90} \subset \mathcal{H}_{0.85}$. Cpd-5, of course, remains an isolated node. In this case, the adjacency matrix simplifies to

$$\mathbf{A}_{0.90} = \begin{pmatrix} \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} & 0 \\ 0 & [0] \end{pmatrix} \tag{1.59}$$

Although the block diagonal structure remains, the main 4×4 block is simpler (i.e., has fewer nonzero elements) than that in Eq. (1.57). In any case, whether a graph-based or matrix-based representation is used, threshold CSNs provide a comprehensive representation of the global "pathways" that connect compounds with respect to a given threshold similarity value. As an example, it is possible to determine the minimum number of edges that must be traversed to go from any given compound to another compound given that the similarities of compounds along the pathway exceeds the similarity threshold value, a feature that can be useful in large screening campaigns but is difficult to carry out in coordinate or cell-based CSs.

As will be seen in the sequel, statistical analyses also play a major role in assessing the characteristic features of networks [129, 177, 178]. In addition, algorithms for treating very large systems such as the Internet as networks has given rise to the development of many powerful methods for handling mega-networks [179]. Thus, representing CSs as CSNs has some distinct advantages as is seen below.

1.3.4.2 Statistical Aspects of CSNs

Vertex Degrees and Degree Distributions Because of their extremely large sizes and complexities, networks are typically characterized in terms of the statistical properties of their vertices and the relationships among subsets of them. One of the most important features of networks illustrated by the simple examples below is *vertex degree*—the number of edges incident on a vertex. [17] The distribution of vertex degrees for large random networks follows a *Poisson distribution* [129] that for networks with very large numbers of vertices becomes

$$\Pr(k) = e^{-\bar{k}} \left(\frac{\bar{k}^{k}}{k!} \right) \tag{1.60}$$

where k is the degree of a randomly chosen vertex and $\bar{k}$ is the mean vertex degree of a large random network. Although it remains finite, for large values of $\bar{k}\,\Pr(k)$ approaches a normal distribution.

It will be seen in the sequel that such networks do not describe typical CSNs. As illustrated in Fig. 1.14a, b, the degree of each vertex in a complete graph is given by $k_i = n-1, \quad i = 1,2,\ldots,n$, where n is the number of vertices in the complete graph; $n = 5$ in the current example. In Fig. 1.14a, $k_i = 5-1 = 4$, while for the complete subgraph $\mathcal{H}_{0.85}$ in Fig. 1.14b, $k_i = 4-1 = 3, \quad i = 1,\ldots,4$, while the vertex degree of the isolated vertex is, of course, zero. In larger, more complex networks, vertex degrees are typically given by statistical distributions as illustrated by the simple example in Fig. 1.14c, where

$$\begin{aligned} k_5 &= 0 \\ k_1 &= k_3 = 1 \\ k_2 &= k_4 = 2 \end{aligned} \tag{1.61}$$

The degree distribution is the probability a given vertex has k incident edges, i.e.,

$$\Pr(k) = \frac{\sum_{i \in k_i = k} k_i}{\sum_{l=1}^{5} k_l}, k = 1,\ldots,5 \tag{1.62}$$

where the term in the numerator is a sum over all vertices of equal degree, and the values corresponding to the example in Fig. 1.14c are

$$\begin{aligned} \Pr(k=0) &= 1/6 \\ \Pr(k=1) &= 2/6 \\ \Pr(k=2) &= 4/6 \end{aligned} \tag{1.63}$$

[17] Although it is not addressed here, the vertex degree of directed graphs/networks can be handled by assessing the “in-degree” and “out-degree” of a vertex that corresponds, respectively, to the number of edges directed towards the vertex and the number directed away from the vertex.

Degree Correlations: Assortativity Coefficients Degree correlations, also called assortativity coefficients, provide a measure of the correlation of vertex degrees between pairs of directly connected vertices. It is obvious from Fig. 1.14a, b that degree correlations for vertices in complete graphs or subgraphs are unity since all vertices in these graphs have identical vertex degrees and hence are maximally correlated. However, the situation in Fig. 1.14c is more complex. The average vertex degree based on the values in Eq. (1.61) is $\bar{k} = \frac{1}{5}(0+1+1+2+2) = 1.2$ and the assortativity coefficients are given by a modified version of the Pearson correlation coefficient [180][18]:

$$\Delta(\mathcal{G}_{0.90}) \doteq \frac{\sum_{i=1}^{5}\sum_{j=i+1}^{5} \mathrm{A}_{0.90}(i,j)\cdot(k_i-\bar{k})\cdot(k_j-\bar{k})}{\sum_{i=1}^{n}(k_i-\bar{k})^2} \tag{1.64}$$

where $\mathcal{G}_{0.90}$ is the threshold graph of $\mathcal{G}$ with respect to a similarity threshold value of 0.90, and $\mathrm{A}_{0.90}(i,j)$ is the *i, j*th element of the adjacency matrix corresponding to that threshold graph. Because of the block structure of the adjacency matrix in Eq. (1.59) only, the vertices corresponding to Cpd-1 through Cpd-4 need be considered in Eq. (1.64).

Carrying out the computation yields a value for the degree correlation of

$$\Delta(\mathcal{G}_{0.90}) = 0.24.$$

Transitivity: Mean Clustering Coefficient Another coefficient of interest is the transitivity or mean clustering coefficient, $\bar{C}(k)$, of all vertices with *k* edges, which can be computed according to:

$$\bar{C}(k) = \frac{1}{N_k}\sum_{i=1}^{N_k} C_i(k) \tag{1.65}$$

where N_k is the number of vertices with *k* edges and $C_i(k)$ is the *local clustering coefficient*

$$C_i(k) = \frac{\varepsilon_i}{\frac{1}{2}k(k-1)} \tag{1.66}$$

with ε_i being the number of edges connecting the *k* neighbors of the *i*th vertex to each other and $\frac{1}{2}k(k-1) = \binom{k}{2}$ is the number of unique pairs of neighbors. Thus, the local clustering coefficient is the ratio of the number of edges connecting the *k* neighbors with each other divided by the total number of *possible* edges among the set of *k* neighbors.

It is clear from Fig. 1.14c that the transitivity in all cases is zero. By contrast, the transitivity of the complete graph in Fig. 1.14a is unity since each vertex has an

[18] Note that the summations are over all unique pairs of vertices (i.e., molecules) and that the coefficient cancels out of the numerator and denominator of Eq. (1.64).

identical number of edges and the vertices connected to that vertex are fully connected with each other, hence, $C_i(k) = 4\cdot 3 \big/ \left[\frac{1}{2}4(4-1)\right] = 1$, which when substituted into Eq. (1.65) gives $\bar{C}(4) = 1$.

Shortest (Geodesic) Path Lengths/Distances In general, a path between vertices can be quite complex as it can include vertices or edges that have been traversed previously. Here, a special kind of path called a *shortest path* is considered. Such paths, also called geodesic paths, are the shortest distance between two vertices based on a count of the number of unlabeled edges in the path. They are not necessarily unique since several paths of equal length may exist in the same graph or network. Shortest path values are entirely equivalent to *graph distances,* $d_{i,j}$, and hence satisfy the well-known distance axioms [177]. A number of algorithms that exist for determining shortest paths have been clearly described in Newman's book [129].

Mathematically, the mean geodesic distance between all unique pairs of vertices is given by

$$\bar{L} = \frac{1}{\frac{1}{2}n(n-1)} \sum_{i=1}^{n} \sum_{j=i}^{n} d_{i,j} \tag{1.67}$$

As can be seen in Fig. 1.14b, the shortest (geodesic) path between two vertices of a complete, unlabeled graph is unity in all cases. This is not the case for the threshold graph in Fig. 1.14c. Computing shortest path lengths in this case is simple since a single path connects the four vertices. Hence, for example, the shortest path between vertex-1 and vertex-4 is of length two and that between vertex-1 and vertex-3 is three. The corresponding mean shortest (geodesic) path length is, from Eq. (1.67),

$$\begin{aligned} \bar{L}(\mathcal{H}_{0.90}) &= \frac{1}{\frac{1}{2}n(n-1)}\left[d_{1,2} + d_{1,3} + d_{1,4} + d_{2,3} + d_{2,4} + d_{3,4}\right] \\ &= \frac{1}{\frac{1}{2}4(4-1)}\left[1+3+2+2+1+1\right] = \frac{1}{6}\left[10\right] \\ &= 1.67 \end{aligned} \tag{1.68}$$

Another feature of shortest (geodesic) paths is of note, namely, they are self-avoiding, as they do not cross themselves. If they did a loop would be formed that could be removed without interrupting the traversal of the path between the specified vertices. Determining shortest paths can be a challenge for large networks, but as noted above, robust path algorithms exist for mega-networks such as the Internet, so dealing with CSs while challenging is not out of the realm of possibility.

Small World Effect The small world effect, namely, that the mean geodesic distance between the vertices in networks defined by Eq. (1.69) is proportional to log *n*, and thus, is generally small for a number of real-world networks (see e.g., Table 8.1 in [129]). A common feature of many small-world and random networks is that their vertex degree distributions tend to be homogeneous with a peak at the mean value of the distribution and an exponential decay, $\Pr(k) \sim \exp(-k)$, in its tail, giving rise to what are called *exponential networks*. Interestingly, there are a number of types of small world networks including ones discussed below that also exhibit scale-free behavior (*vide infra*) [181].

One consequence of the small world effect is the famous "six degrees of separation" hypothesis, namely, that everyone on Earth is separated by no more than five individuals (vertices) and hence six links (edges). That this is not an entirely unreasonable hypothesis is based on the following overly simplistic argument. Suppose I have 100 friends each of which has 100 friends, each of which has 100 friends, etc. Thus, with only one degree of separation I can connect to 100 individuals, with two degrees I can connect to $100 \times 100 = 10,000$ individuals, and with only three degrees of separation I can connect to $100 \times 100 \times 100 = 1,000,000$ individuals. If all six degrees of separation are considered, I could potentially connect to one trillion individuals, 50 times more than required to connect to everyone on Earth. Although, as pointed out by Watts [127] this argument has significant practical flaws, it nonetheless captures some essential features of small-world networks.

Networks exhibiting small-world behavior, hence, can facilitate many processes such as communication, the spread of disease, and the speed of inter-server access on the Internet. Not surprisingly, as will be discussed in Sect. 1.3.5.2, CSNs tend to exhibit small world behavior as well. This is not surprising given the nature of molecular and chemical similarity, which in general does not exhibit transitive behavior: i.e., if A is similar to B and B is similar to C, it does not in all cases follow that A is similar to C. This same phenomenon exists in social networks as well, i.e., if A knows B and B knows C it does not mean that A and C also know each other, although the likelihood that they do is higher than random chance. As discussed by Newman [129], transitivity is related to various forms of clustering coefficients.

Scale-Free Networks The vertex degree distributions of scale-free networks differ from those of large random networks and many small world networks, which are Poisson distributed (*vide supra*). By contrast, scale-free networks described by Barabási and Albert [182] are nonhomogeneously distributed and follow power laws, such that the probability that a random vertex has degree k [19] is inversely related to a power of vertex degree, i.e.,

$$\Pr(k) = \kappa \cdot \left(\frac{1}{k}\right)^{\alpha} = \kappa \cdot k^{-\alpha} \tag{1.69}$$

where κ is a constant and the exponent $\alpha > 1$ is a *scaling coefficient,* which usually lies in the range $2 \le \alpha \le 3$ for many real-world networks (see e.g., Table 8.1 in [129]). Van Steen gives a clear description of why the power law given by Eq. (1.69) is scale-free [180]. In addition, the mean shortest path length of scale-free networks is proportional to log log *n,* a value that is much less than the log *n* behavior noted above for many small world networks.

Two important properties of scale-free distributions are that they do not have peaks and they decay at much slower rates than the corresponding Poisson and

[19] Note that this can also be interpreted as the fraction of vertices of degree k.

normal distributions. The second property is especially important because it indicates a higher probability that more extreme events may occur than can occur in the latter distributions. In this regard, an important example in the case of scale-free networks is the presence of vertices with exceptionally high vertex degrees, a situation that gives rise to highly connected "hubs" interconnected by relatively small numbers of edges, a rather extreme form of small world behavior to say the least.

Because of its form, depicting Eq. (1.69) as a $\log \Pr(k) \text{versus} \log k$ plot should result in a straight line if the distribution does follow a power law, at least asymptotically. Proving that it does is not necessarily easy, since some values of k in the tail of the distribution may not satisfy the power law relationship. However, as pointed out by Newman among others [129], alternatives exist that provide a means for accomplishing this, although sometimes it requires removing some of vertex degrees that do not follow the power law.

1.3.4.3 Topologies of CSN

As noted earlier, five papers have been published that address various aspects of *similarity-based networks* of CSs [168–172], all of which differ from the related work on power laws in CSs by Benz et al. [173] that predates these papers. Both of the latter reports have presented evidence of the small world behavior of CSNs and in some cases scale-free behavior as well. Because the edges of the CSNs are unlabeled, threshold graphs were generated for different similarity threshold values. Not surprisingly, statistical features related to vertex degree tend to decrease as the similarity threshold is raised as is nicely illustrated in Table 1.2 of reference [169].

Although this behavior seems intuitive, it can be rationalized as follows. Due to the central limit theorem [183], the set of similarity values associated with large compound DBs is normally distributed with a mean around, say for example, 0.50. Now arrange the set of similarity values in *descending* order and determine the corresponding *cumulative probability distribution* depicted in Fig. 1.15, where the abscissa corresponds to the threshold similarity value for a given CSN, and the ordinate corresponds to the *fraction,* f_{edge}, of the $n(n-1)/2$ possible edges that can be drawn between the n compounds that constitute the vertices of the network. It is clear from the figure that for a threshold similarity value of 0.75 less than 10 % of the compounds will be connected directly. Even at a threshold similarity of 0.5 only about half the possible number of edges are present.[20] In order to gain a sense of the magnitude of the problem, consider a DB of only $n = 10{,}000$ compounds. In this case, the *complete* CSN would have ~ 50 million edges. However, even at a similarity threshold value of 0.75 about 8 % of the total possible edges (~4,000,000 edges) will be formed. As this is more than 400 times the minimal number of edges needed to connect all of the vertices with one another (~ 10,000), it is certainly sufficient to introduce significant and interesting structure in the CSN. Hence, it easy to see

[20] This argument is, of course, oversimplified since it depends on the width (standard deviation) of the probability distribution.

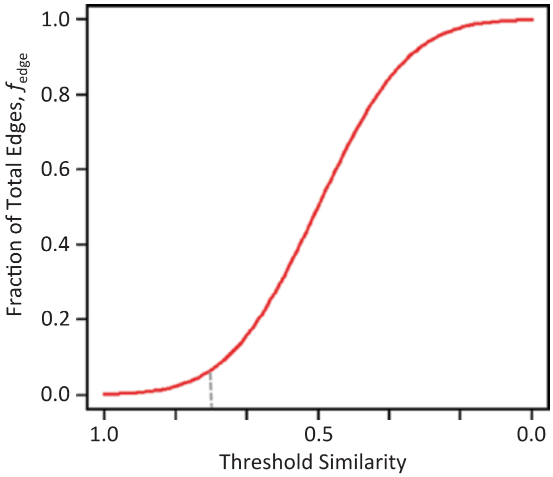


Fig. 1.15 Cumulative distribution curve showing the fraction of possible edges formed as a function of similarity threshold value. The light grey dashed line corresponds to a threshold similarity value of 0.75

that expanding to a DB of say 200,000 compounds can prove to be a challenging enterprise.

The paper by Tanaka et al. [168] investigates small world phenomena in several libraries obtained directly from the ZINC DB [184] and from virtual libraries constructed from structurally diverse fragments. By contrast, the paper of Krein and Sukumar [169] undertakes a much more comprehensive analysis based on a number of different sets of CS descriptors applied not only to CSs but also to their subspaces associated with activity cliffs. A recent paper from Bajorath's group [172] also addresses subnetworks associated with activity cliffs. Obviously, these analyses can be extended to other landscape features such as similarity cliffs (see Sect. 1.2.4).

The approximately scale-free nature of CSNs observed by Krein and Sukumar led them to infer the existence of hubs, highly interconnected regions of CSNs linked together by relatively sparse paths. Hubs represent regions of CS associated with different structural motifs. Hence, paths linking hubs may provide a means for addressing the problem of scaffold hopping, a process associated with the presence of similarity cliffs, which are more general since they include scaffold hops as a special case.

Another application of threshold CSNs is exemplified by the work of Bajorath's group on network-like similarity graphs (NSGs). NSGs are threshold graphs they developed as a means for analyzing the SARs of large, diverse sets of compounds. Figure 1.16 provides an example of an NSG that characterizes the activities of a set of lipoxygenase inhibitors taken from the paper by Wawer et al. [170]. Compound potencies are color coded from red for the most active (1 nM) to green for the least active (100 μM). Links are drawn between compound pairs if their MACCS Tanimoto similarity exceeds 0.65. Additional annotation corresponds to SAR index scores (decimal values) associated with compound clusters. The index ranges from 0.00 to 1.00, the larger the value the more "discontinuous" a given compound cluster—activity cliffs correspond to high levels of discontinuity.

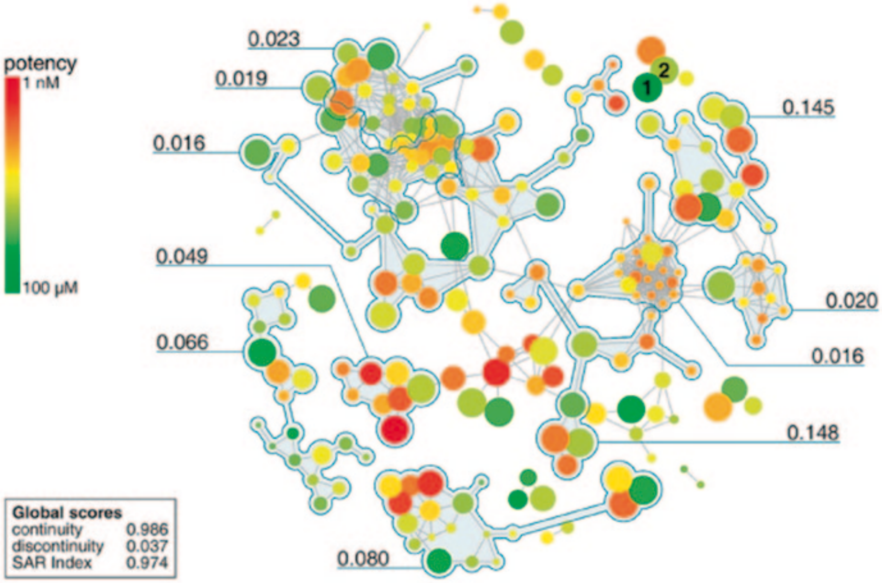


Fig. 1.16 Network-like similarity graph (NSG) depicting the CS and activity relationships of a set of lipoxygenase inhibitors taken from the work of Wawer et al. [170]. Compound potencies are color coded as shown by the colored bar on the upper left hand side of the figure, red being the most active and green being the least active. Compounds are connected by an edge if the MACCS Tanimoto similarity value of a given compound pair exceeds 0.65. The *decimal numbers* associated with clusters of compounds correspond to SAR Index scores (See text for additional details)

1.3.5 Exploring CSs

The concepts of structural similarity and CS, which are ubiquitous in medicinal chemistry, are finding a place in other chemically related sciences such as materials science and engineering [185]. A question that now arises is how can we develop procedures and algorithms that exploit these concepts to facilitate the discovery of new drugs and bioactive agents? Or, more appropriate to the book in which this chapter resides, how can these concepts be applied in food science and in aroma and flavor chemistry? Although the examples presented in this section do not represent a comprehensive set of the many possible methods that are available, they will at least provide a sample that should afford sufficient information to help answer this question.

1.3.5.1 Comparing Compound DBs

It is obvious from previous discussion in this chapter that compound DBs play an extremely important role in many aspects of chemical informatics. Thus, it is important that methods exist for assessing their similarities and differences. As has been noted by a number of investigators cell-based methods are particularly suited to this task.

For example, consider the compound DBs listed in Table 1.4 and discussed in Sect. 1.3.3.2. While the numerical values in the table provide a reasonable summary of the cell-based characteristics of each collection, they are not specific enough to afford a detailed *comparative* assessment, as they do not account for relationships between the cells in collections being compared. Pearlman and Smith [76] developed an approach that is able to address this deficiency, albeit only partially.

The procedure is as follows. First, a cell occupancy threshold is chosen; in the example discussed here, an occupancy value ≥ 1 is used, i.e., each occupied cell contains at least one compound. Obviously this is a potential source of error since an occupied cell in one collection could contain a single compound, while the corresponding cell in another collection could be occupied by, say, more than a 100 compounds. Hence, the Pearlman–Smith (P–S) procedure only compares patterns of occupancy, but this may be sufficient when very large compound collections of comparable size are being compared, or if only a coarse-grained estimate is required. Carrying out the analysis for a sequence of occupancy thresholds, e.g., $t_{\text{occ}} \geq 1, \geq 2, \geq 3, \ldots$, would provide a measure of the sensitivity of the results to the chosen occupancy threshold, but such an approach to my knowledge has not been carried out.

The P–S procedure can be viewed in a manner that is entirely equivalent to that described earlier for binary FPs since the set of cells in a cell-based CS can be thought of as one long FP. How the cell-based CS is unfolded into the linear array of cells is unimportant; what is important is that all equivalent cell-based CSs that are compared be unfolded in exactly the same way. Occupied cells are labeled with a “1” if they are occupied by at least one compound and by a “0” if they are unoccupied. Hence, any of the FP-based similarity coefficients can now be used to assess the similarity of any pair of compound collections or libraries described by the same cell-based CS. These “DB FPs” are on the order of 100,000 or more cells, and hence, many times larger than typical binary structural FPs that usually have less than 2000 elements. And, as seen in Table 1.4, only a small fraction of the cells are occupied so that these FPs are very sparse. The discussion in Sect. 1.2.1.1 shows that they can be handled using run-length encoding, or a similar procedure. Additional compression, such as is the case for some large molecular FPs, is not necessary in this case since the number of DBs being compared is many times smaller than the number of molecular FPs typically dealt with in similarity search-based activities.

The P–S procedure defines two measures for assessing the similarity of two compound DBs, nominally A and B, residing in the same CS:

$$\begin{aligned}&1.\text{Fraction of A}'\text{s cells occupied by B}\\&2.\text{Fraction of B}'\text{s cells occupied by A}\end{aligned} \tag{1.70}$$

These definitions are completely equivalent to the *asymmetric* Tversky measures given in Eqs. (1.12) and (1.13), respectively, and can be interpreted in a like manner,

Table 1.5 Comparison of percent occupancies of compound collections in six-dimensional 3-D BCUT chemical space based on the P–S procedure

A\B	Diverse	Combi	MDDR	Micros
Diverse		11.7	43.8	2.8
Combi	88.5		85.2	6.0
MDDR	78.9	20.3		44.0
Micros	98.5	28.3	86.6	

See Table 1.4 for details of compound collections. Cell occupancies ≥1

but any of the similarity coefficients described in this work that are based on binary structural FPs can be used. Note that the two expressions given in Eq. (1.70) can also be interpreted probabilistically.

Since the set of cells in a cell-based CS are analogous to binary structural FPs, other similarity measures such as those based on the Tanimoto or Dice similarity coefficients given in Eqs. (1.8) and (1.9) can be used. Alternatively, the corresponding dissimilarity coefficients given in Eqs. (1.21) and (1.22) also can be used. As noted in Sect. 1.2.1.3, the numerator of the Tanimoto dissimilarity coefficient is just the Hamming distance, which is a measure of the number of differences between the two DB FPs.

Table 1.5 provides an example of how the similarity measures given in Eq. (1.70) can be applied to a more detailed assessment of the similarity of pairs compound collections. For example, 0.885 of the occupied cells in the Combi collection are also occupied in the Diverse collection. Conversely, only 0.117 of the occupied cells in the Diverse collection are also occupied in the Combi collection, a clear example of the much greater diversity inherent in the Diverse collection. In contrast, 0.985 of the occupied cells in the Micros collection are also occupied by the Diverse collection, while only 0.028 of the occupied cells in the Diverse collection are also occupied in the Micros collection—not a surprising result given that only 516 cells are occupied by the entire Micros collection. Thus, although in relative terms the Micros collection is diverse, in absolute terms it does not compare with that of the Diverse collection.

1.3.5.2 Subset Selection and Compound Acquisition

Subset Selection Subset selection is used primarily for assembling diverse subsets of compounds for HTS campaigns. Another form of subset selection called *similarity searching* or LBVS also requires activity data, albeit on a small subset of compounds, as will be discussed in Sect. 1.3.5.4. Hence, subset selection usually takes places in early screening while similarity searching or LBVS is typically used in subsequent follow-on screening activities. Because in the former case activity data are generally unavailable, constructing appropriate subsets of compounds for the initial phases of an HTS campaign can be challenging [186–189].

While there are many variations, the underlying strategy for generating initial screening sets almost always relies on maximizing their diversity by minimizing the similarity (or maximizing the dissimilarity) of the compounds in the putative screening set. It is important to note that unlike similarity or dissimilarity, which are pairwise measures, diversity is a population-based measure associated with the dissimilarity of the entire subset of compounds [10, 41]. In this regard, a number of authors have addressed the issue of how to estimate the diversity of a large collection of compounds [190–192]. Willett [193, 194] and Agrafiotis [191] have presented descriptions of many aspects of diversity-related methods and procedures. An interesting discussion of the early history of the concept of molecular diversity was published in 2001 [195].

Although the field of molecular diversity is vast, the focus in this work is on two approaches: on cell-based sampling of CS [76] and on a maximum dissimilarity/distance algorithm called "Dfragall" [63]. Here the terminology MaxD will be used in place of Dfragall to indicate the generality of the procedure. Both approaches generally use 2-D structural information, although the use of 3-D BCUTS does account, albeit in a somewhat limited fashion, for 3-D information. Matter has presented a more detailed comparison of the role of 2-D and 3-D descriptors in selecting diverse subsets of compounds [196]. As will be seen in the following subsection on compound acquisition, the cell-based approach is clearly superior in its ability to identify and fill so-called "diversity voids," which can be important in a number of instances.

A variety of cell-based sampling schemes can be employed in order to obtain a subset of the desired size and diversity [76, 78]. These schemes include *simple sampling*, where a single compound is obtained from each occupied cell, *threshold-based sampling*, where the number of compounds selected from each cell is less than (if the cell has fewer compounds than the threshold value) or equal to the threshold value, *proportional sampling*, where the size of the sample is proportional to the number of compounds in the cell, or *property-based sampling*, where compounds are selected based on a range of values for one or more properties such as molecular weight or log*P*. Property-based sampling can, of course, be applied simultaneously with any of the other sampling procedures. If the size of the desired sample is less than the number of compounds obtained by a given sampling procedure, either fewer cells can be sampled or the number of compounds per cell can be reduced. In the former case, since neighborhood relations among cells are not considered in cell-based CSs, a random selection of sampled cells could be considered. By contrast, the subset selection procedure based on MaxD is much more computationally demanding and does not explicitly fill diversity voids, although it may inadvertently do so to some degree. In the MaxD case, a typical selection procedure is shown in Table 1.6.

An example that illustrates, but of course does not generally prove, the superior performance of cell-based compared to dissimilarity-based subset selection is depicted in Fig. 1.17. The computations were carried out in 3-D BCUT CS based on the Diverse DB (see Table 1.4) described earlier. The cyan dots in the 2-D projection of the CS depicted in Fig. 1.17a, b represent the compounds in the DB, while the yellow dots represent the compounds obtained in each of the sampling procedures. In the MaxD subset selection depicted in Fig. 1.17a, only about 36 % of the

Table 1.6 MaxD subset selection procedure

Step	Procedure
1	Choose a compound, x_1, at random from the compound collection of interest
2	Determine x_2, the compound most dissimilar to or most distant from x_1
3	Determine x_3, most dissimilar to or distant from compounds x_1 and x_2
4	Repeat the process until the desired number of compounds is obtained or the chosen dissimilarity or distance value falls below the chosen threshold value or reaches a plateau

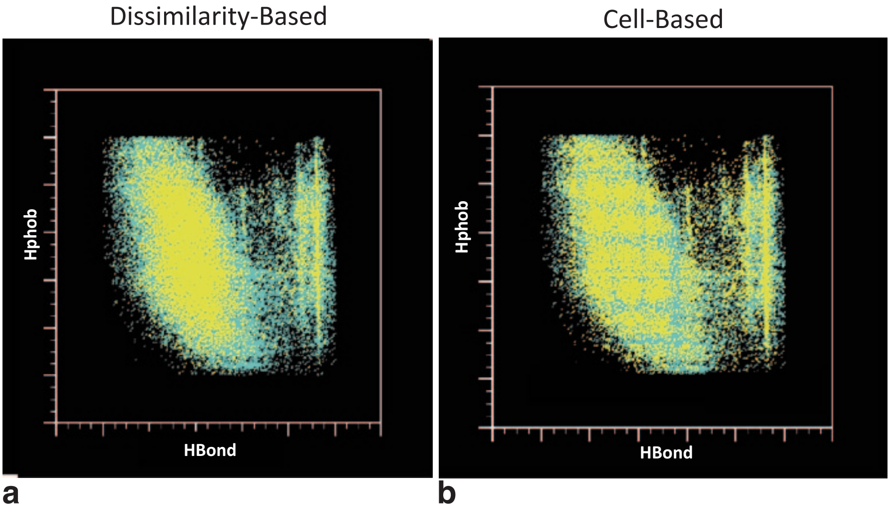


Fig. 1.17 Comparison of subset selection procedures based on compounds in the Diverse collection depicted in cyan (see Table 1.4 and Sect. 3.6.1 for details). Yellow dots represent compounds obtained by the subset selection procedures: **a** dissimilarity-based selection. **b** Cell-based subset selection. (Figure kindly provided by Veer Shanmugasundaram)

original 18,371 occupied cells in the associated cell-based CS are occupied by at least one sampled compound. By contrast, 100 % of the available cells are occupied in the cell-based procedure by a similar number of compounds to that obtained by the MaxD algorithm, which is not surprising since the cell-based procedure is based on sampling each cell of the CS. This affirms, but certainly does not prove, what is intuitively expected, namely, that the cell-based procedure results in broader sampling than the corresponding MaxD procedure.

Compound Acquisition There are two general goals associated with compound acquisition—enhancing the *diversity* of an existing collection and maintaining its *integrity*. While the focus is generally on the former, the latter is also important due to the rate at which compounds can be used up in assays and related activities or can decompose over time. Enhancing diversity usually involves filling unoccupied or partially occupied regions of CS. Maintaining DB integrity, on the other

Table 1.7 Compound acquisition procedure

Step	Procedure
1	Identify vendor collections from which to purchase compounds and preprocess them to remove “undesirable” compounds
2	Generate a cell-based chemical space containing the combined original compound DB and appropriate vendor DBs
3	Select the initial set of vendor compounds by filling diversity voids
4	Additional diversity assessment of the initially selected set of vendor compounds using a modified MaxD algorithm (see Table 1.8)
5	Apply compound filters that were developed based on the knowledge of experienced medicinal chemists
6	Direct review by medicinal chemists
7	Submit compounds for purchase

hand, involves replenishing DB compounds that have become depleted or, if exact replacements are unavailable providing compounds that are, at least to some degree, similar to the original ones. A number of papers addressing compound acquisition have been published over the years, a sampling of which is given by the following references [162, 197–199].

The following is a brief description of the acquisition process based on the work reported in [162]. It illustrates a number of the general issues that must be dealt with, but since there are many ways to do so, what is given here should only be considered a rough outline of an acquisition process. The papers just cited should be consulted for additional examples. Table 1.7 provides a summary of the compound acquisition procedure.

A number of issues arise in step-1, especially when the purchase of large sets of compounds is desired. Some of which include the presence of compounds with undesirable features (e.g., nitro groups) in a vendor’s collection and whether the compounds are “Lipinski compliant,” i.e., obey the rule of five [200]. Although the rule of five was intended primarily to address potential drug delivery and bioavailability issues, it has become a surrogate for drug likeness, and its application has far exceeded the developers’ initial intentions as to its domain of applicability. A recent procedure suggests a modification of the rule of five that increases its robustness to small differences in the parameter values, although it does not extend its domain of applicability [201]. In a related study, Bickerton et al. [202] developed a similar, but more comprehensive procedure that takes account of additional features, namely, molecular polar surface area, number of rotatable bonds, number of aromatic rings, and number of structural alerts, typically associated with drug likeness. In addition, diversity and structural novelty of a collection, timely availability of compounds, and compound purity are other desirable characteristics of vendor compound collections.

In step-2, there are several choices of methods to carry out the initial selection of compounds. The cell-based approach is employed here because of its computational speed and ease of application. Figure 1.18 depicts a model of a cell-based sampling scheme similar, but not algorithmically identical, to that implemented in *Diverse Solutions*™ [78] (cf. [63]) and presented in a way that is designed to clarify the

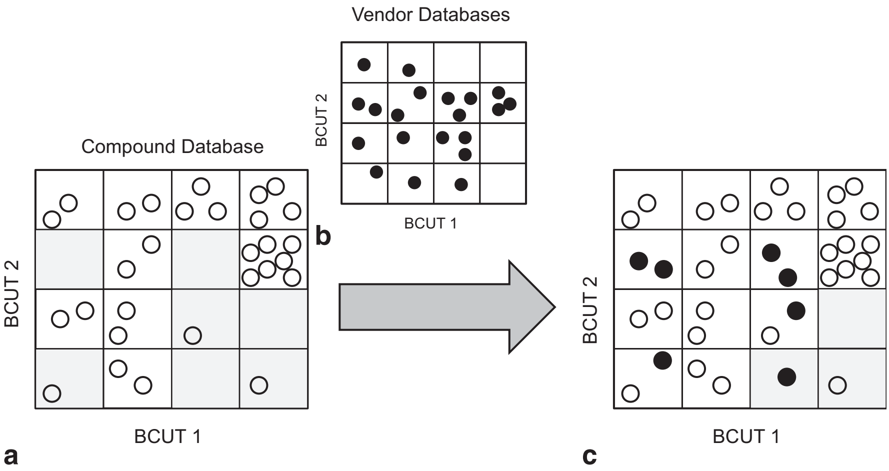


Fig. 1.18 Schematic depiction of a model 2-D cell-based selection process for compound acquisition (Cf. [162]). In **a** *unfilled circles* represent compounds in the original compound DB; in **b** *filled circles* represent compounds in the combined, pre-processed vendor DB; **c** depicts the augmented compound DB after the initial selection process has been completed. Cells shaded in *light grey* represent diversity voids for cells containing fewer than two compounds. (See text for addition details)

compound selection process. A two-dimensional BCUT CS is generated by *combining* (using set-theoretic union) the set of compounds in the original compound DB, O^{DB}, and the compounds in the set of vendor DBs $\mathrm{V}^{\mathrm{DB}} = \left\{\mathrm{V}_1^{\mathrm{DB}}, \mathrm{V}_2^{\mathrm{DB}}, \mathrm{V}_3^{\mathrm{DB}}, \ldots\right\}$, where $\mathrm{V}_i^{\mathrm{DB}}$ is the set of compounds in the *i*th preprocessed vendor DB:

$$\begin{aligned}\widehat{\mathrm{M}} &= \mathrm{O}^{\mathrm{DB}} \cup \mathrm{V}_1^{\mathrm{DB}} \cup \mathrm{V}_2^{\mathrm{DB}} \cup \mathrm{V}_3^{\mathrm{DB}} \cup \cdots \\ &= \mathrm{O}^{\mathrm{DB}} \cup \mathrm{V}^{\mathrm{DB}}\end{aligned} \tag{1.71}$$

$\widehat{\mathrm{M}}$ is then used as a basis for constructing a CS that includes all of the original and preprocessed vendor compounds, which can be written symbolically as $\widehat{\mathrm{M}} \Rightarrow \mathrm{CS}(\widehat{\mathrm{M}})$.

Figure 1.18a shows the distribution of the original set of compounds in the newly constructed CS. Likewise, Fig. 1.18b shows the distribution of the vendor compounds in the same CS. In the cell-based approach, empty cells as well as those with very few compounds, say less than two or three, can be considered to be diversity voids. Such cells are suitable candidates for compound acquisition. In the example in Fig. 1.18a, there are four empty cells and three cells containing single compounds, all shaded in light grey, which can be classified as diversity voids in this model DB. Now compounds from the combined vendor DB depicted in Fig. 1.18b are used to fill the diversity voids in in Fig. 1.18a until the cell occupancy of all cells in the DB is at least two. This is illustrated in Fig. 1.18c, where the cells shaded in light gray indicate diversity voids that remain after compound acquisition. As seen in the figure, some of the empty cells are now populated with vendor's compounds

Table 1.8 Diversity assessment using a modified MaxD subset selection procedure

Step	Procedure
1	Determine vendor compound, x_1, that is most dissimilar to all of the compounds in the original compound database (C-DB) and add it C-DB giving C-DB + x_1
2	Determine the vendor compound, x_2, that is most dissimilar to C-DB + x_1 and add it yielding C-DB + x_1 + x_2
3	Repeat steps 1 and 2 until the desired number of compounds is obtained or until the dissimilarity value falls below a specified threshold

and some remain unoccupied, as no vendor compounds existed for those cells. The third cell from the left in the bottom row of Fig. 1.18c, which was unoccupied originally, is now occupied by a single vendor compound since only one such compound was available to fill that cell as seen in Fig. 1.18b.

The basic idea here is to populate unpopulated cells and those of low occupancy with commercially acquired compounds. As was the case in subset selection, there are a number of ways in which cells can be populated with new compounds, the simplest being to populate all unpopulated cells with at least one compound. While such an approach is straightforward, it is not, in general, a practical strategy. An examination of Table 1.4 clearly shows why this is the case. In that example, the 6-D CS contains 117,649 cells, 18,371 of which are occupied by at least one compound. This leaves 99,278 empty cells. Even if a set of sufficiently diverse compounds were available for purchase the cost would be significant—at an average price of $ 25 per sample, this would amount to nearly $ 2.5 million, an amount that would test the budget of all but the largest pharmaceutical companies. Thus, additional strategies need to be implemented to address compound acquisition in a way that ensures an optimal, albeit incomplete, selection is made [162].

Although the number of cells in cell-based CS is large, the hyper-dimensional volume of each of the cells is also large. Hence, compounds within a given cell may be quite dissimilar. In contrast, compounds located near a common boundary between two cells may be quite similar even though they reside in different cells (*vide supra*). Because of this type of "idiosyncratic" behavior associated with cell-based CSs, and additional level of similarity analysis may be warranted to ensure that the selected compounds are as dissimilar to each other as possible. This can be accomplished in step-4 using a modified form of the MaxD ("Dfragall") algorithm [63] based on Euclidean distance computed with respect to the BCUT coordinates or, as is traditionally done in the algorithm, using some form of similarity/dissimilarity measure, a procedure that further reduces the number of compounds.

An alternative approach to that described above has been described by Lajiness [63]. It is a variant of the MaxD ("Dfragall") algorithm presented earlier and is summarized in Table 1.8. One clear deficiency of this algorithm is that it is difficult to fill specific diversity voids.

In step-5 of Table 1.7, a set of compound filters based on the knowledge of experienced medicinal chemists is applied further reducing the size of the set of potential compounds for acquisition. Examples of these filters include a number of compound characteristics such as number of rings (2–4), molecular weight (200–400), number of rotatable bonds (0–5), log*P* (−1 to 2). Finally, in step-6, medicinal chemists directly evaluate the remaining molecules [116], and those that survive this final review are submitted for purchase.

1.3.5.3 Similarity Searching and LBVS

Basically, there are three in silico approaches used to the identify compounds with potential biological activity all of which fall under the rubric of virtual screening methods:

- Ligand–protein docking
- Similarity searching based on 2-D molecular descriptors (2-D LBVS)
- Similarity searching based on 3-D molecular descriptors (3-D LBVS)

A number of edited volumes [164, 203–205] and reviews [104, 206–215] have addressed many aspects of virtual screening; and Parker and Bajorath have discussed an important but rarely touched upon issue concerning the effect of errors on both HTS and LBVS [216].

Ligand–Protein Docking[21] Docking involves two basic steps, finding an optimal structure of the ligand–protein complex and scoring, in some fashion, the fitness of that complex. An advantage of this approach is that it does not require any prior knowledge of biological activity. On the other hand, it does require knowledge of the 3-D structure of the target protein, or of some closely related protein that can serve as a model of the desired target protein, to which the ligand can be docked. However, this is just the tip of the iceberg, as there are many complex issues that must be dealt with in ligand–protein docking including protein flexibility, ligand sampling, and effective scoring functions. In addition, if biological activity requires specific changes in protein structure induced by ligand binding and/or if the solution environment plays a crucial role in the functioning of the protein, then these added complications must also be addressed. And there are other factors some known and some unknown that can further complicate the docking process [217–219].

Similarity Searching There are two types of similarity searching procedures—also called LBVS—that are classified according to the dimensionality of their feature descriptors. 2-D methods employ structural FPs or vector-based descriptors as described in Sects. 1.2.1 and 1.2.2, while the corresponding 3-D methods involve matching pharmacophores [153, 220–223] or molecular shapes [224–226]. Since 3-D methods appear to contain more structural information such as stereochemistry, which in many cases is important for activity, it is surprising that 2-D methods tend to outperform or at least perform comparably to 3-D methods. There are

[21] There are, of course, other docking processes that are of importance in biology including protein–protein, ligand–nucleic acid, nucleic acid–nucleic acid docking to name a few. Ligand–protein docking is highlighted in this work because of its importance in drug discovery and its widespread application in that field.

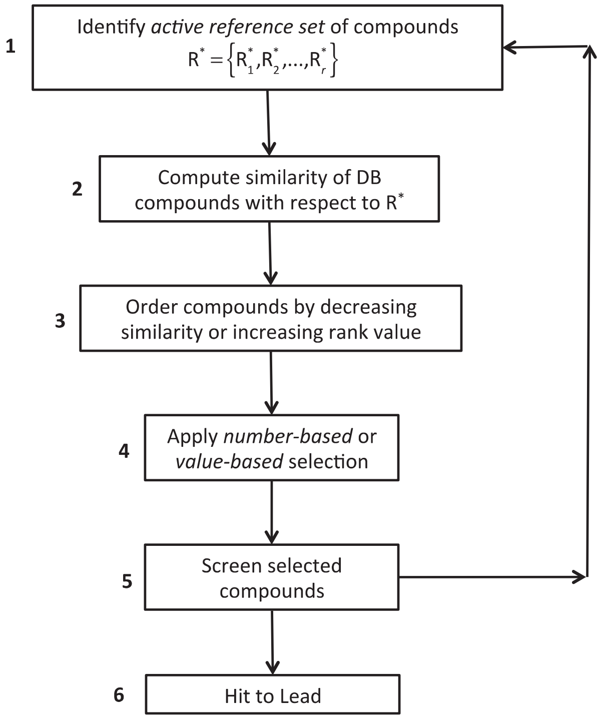

1
Identify *active reference set* of compounds
$R^* = \{R_1^*, R_2^*, \ldots, R_r^*\}$
2
Compute similarity of DB compounds with respect to R^*
3
Order compounds by decreasing similarity or increasing rank value
4
Apply *number-based* or *value-based* selection
5
Screen selected compounds
6
Hit to Lead

Fig. 1.19 Ligand-base virtual screening procedure

many possible reasons for this observation including the fact that the topological structure encoded in 2-D representations may more than compensate for missing 3-D information [10, 18, 88, 227, 228]. In addition, determining the ensemble of biological active conformations can be a difficult and uncertain task [229], and the many approximations made to increase computational efficiency and reduce computing time, also contribute to the somewhat problematic performance of 3-D-based approaches. Hence, in keeping with the discussion in the rest of this chapter, the focus here is on the simpler and faster 2-D LBVS methods.

2-D LBVS[22] Although Stanton et al. [230] were, perhaps, the first group to explore the application of similarity-based techniques in HTS, many examples of LBVS have been published since then, especially in the first decade of the twenty-first century as can be seen from the following references [32, 33, 86, 104, 231–233] and those cited at the beginning of Sect. 1.3.5.3.

As depicted in Fig. 1.19, LBVS is typically an *iterative process*. In step-1, an *active reference set* of compounds is identified in some manner, usually in an HTS campaign. In step-2, the similarity values with respect to each of the actives in R^* are computed. Several cases arise in this regard. First, consider the simplest case of a single active reference compound, which may obtain in many instances, at least

[22] See Sect. 1.2.3 for related discussion.

in the initial iteration of the LBVS process. The compounds are then arranged in decreasing order of their similarity values, or in ascending order by their ranks, one being the highest rank. If, on the other hand, a distance-based measure of similarity is used, the list of compounds will be ordered from smallest distance to the largest distance value. The rank ordering will remain the same, one again being the highest rank. A subset of the top-"scoring" compounds (i.e., compounds with the largest similarity or smallest rank values) is selected. This can be accomplished in two ways, *number based* or *value based*. In the former case, a number of compounds, say the top 100, are selected for follow-on screening regardless of their similarity values or rankings, whereas in the latter, a subset of compounds all of whose similarity values or rankings with respect to R^* are less than or greater than, their respective threshold similarity or ranking values. Regardless of how the compounds are selected, they are screened yielding a new set of actives, and the process is repeated.

This, however, raises a new issue, namely, how are multiple active reference compounds handled in the LBVS process? There are several approaches to this problem. One way is through the use of group fusion described in Sect. 1.2.3.2, which is ideally suited to deal with this problem since multiple active reference compounds are an inherent feature of the method. And, as discussed in Sects. 1.2.3.2 and 1.2.4, group fusion exhibits excellent performance as a means for identifying new actives. Interestingly, group fusion based on the fusion maximum similarity or minimum distance values is essentially identical to an approach called list-based searching [76, 78, 86].

This completes step-3 regardless of whether singleton or multiple active reference compounds were dealt with in that step. Obtaining a subset of the compounds from the resultant ordered list using either number- or value-based selection then completes step-4. In step-5, the resulting set of compounds is then screened. At this point, a choice must be made. If, after screening is completed, it is determined that a sufficient number active compounds of appropriate quality have been obtained, the process may then move to step-6 where the hit-to-lead phase of the drug discovery process can commence, otherwise the process moves back to step-1 and the process is repeated. It is well to note that identifying active reference sets may also include additional assays designed to more firmly establish the biological or pharmacological characteristics of the compounds, and thus to help in determining whether compounds active in HTS should be considered further.

Aggregating the Results of Individual Similarity Searches As discussed in Sect. 1.2.3, combining ("fusing") similarity values, which falls within the class of data aggregation methods [97], has been shown to yield improved results in similarity searches. Generally, fusion methods combine similarity (distance) values or rankings to yield new fused values prior to any similarity search. An alternative approach is to carry out multiple similarity searches on the same set of active reference compounds using different similarity or distance measures and then combine the sets of compounds obtained in this way [86], employing what can be called *post-search aggregation* (PSA). Although related, this differs from similarity fusion that, as discussed in Sect. 1.2.5.1, combines the similarity values and then carries out a similarity search using the fused values.

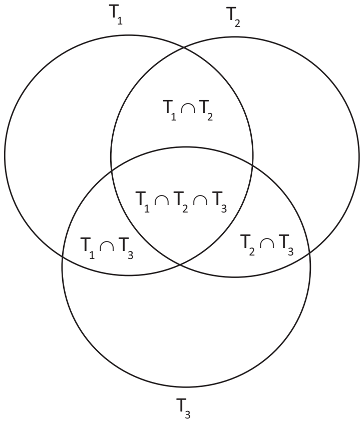


Fig. 1.20 Venn diagram representing the possible joint subsets obtained from three sets of compounds T_1, T_2, and T_3 retrieved by three different similarity or distance-based search methods of a compound DB

A difficulty with PSA methods is that the subset of compounds retrieved in each of the similarity- or distance-based searches may differ significantly. As an example, consider the family of three subsets of compounds retrieved by three corresponding similarity or distance-based searches of a compound DB, i.e.,

$$T = \{T_1, T_2, T_3\} \tag{1.72}$$

where the size of each of the subsets may be taken to be the same and can be determined by a number- or value-based procedure, or the sizes can, if desired, all be different. It is possible and, in fact, occurs frequently that some compounds may be found in more than one of the subsets. The Venn diagram depicted in Fig. 1.20 indicates this. As will be seen in Eq. (1.73), the smaller the "overlap" among the subsets, as measured by set intersection, the broader the sampling of the CS represented in a compound DB.

The basic assumption underlying this approach is that multiple searches using different similarity or distance measures will give rise to higher enrichment factors in a common assay than would be obtained using a single search method. To see this, consider the *background enrichment factor* for a given assay, $E_{\text{Background}}$, which is basically the *estimated* fraction of active compounds in a DB, an estimate usually arrived at by the assay of compounds randomly selected from the DB.

When considering all three subsets, the *breadth or diversity* of the search can be defined as

$$\Delta = \frac{\text{Card}(T_1 \cup T_2 \cup T_3}{\text{Card}(T_1) + \text{Card}(T_2) + \text{Card}(T_3)} \tag{1.73}$$

which satisfies $0 \le \Delta \le 1$, where "Card" refers to the cardinality (i.e. number of elements) in a given set (see also footnote *a* in Table 1.1). The union of the three subsets is the set of compounds unique to all three subsets. Similar expressions can be constructed for the pairwise case by removing the extraneous subset(s).

The singleton case is trivial since $\Delta = 1$. As can be seen from Eq. (1.73), as the breadth approaches unity, i.e., as $\Delta \to 1$, the sampling of CS increases reaching a maximum at unity. However, this procedure is of real value only if it leads to enhanced enrichment factors. The enrichment factor for the three sets of retrieved compounds can be obtained as follows:

The fraction of actives obtained from the three samples is given by

$$f_{\text{sample}} = \frac{\text{Card}\left(\text{T}_1^* \cup \text{T}_2^* \cup \text{T}_3^*\right)}{\text{Card}\left(\text{T}_1 \cup \text{T}_2 \cup \text{T}_3\right)} \tag{1.74}$$

where the asterisks in the numerator denote subsets of actives, such that $\text{T}_i^* \subseteq \text{T}_i$ for $i = 1, 2, 3$ and 'Card' refers to the cardinality, that is the number of elements in the sets. The *enrichment factor* is then given by

$$EF = \frac{f_{\text{sample}}}{f_{\text{background}}} \tag{1.75}$$

where $f_{\text{background}}$ is the fraction of actives obtained from a random sampling of the compound collection of interest.

Interestingly, the procedure appears to be a combination of group fusion (i.e., list-based searching) and similarity fusion. The reasons, the first two of which are associated with group and similarity fusion, are as follows: (1) multiple active reference compounds are used, (2) the most similar (closest) compounds to each active reference compound are retained, and (3) multiple similarity measures are applied.

This approach was described in Shanmugasundaram et al. [86], who investigated its application to a number of targets including those associated with anxiety, Alzheimer's disease, and pathogenic bacteria. The data provided below are based on a bacterial enzyme target and a set of 12 well-characterized active reference compounds. A distance measure based on three different sets of BCUT descriptors and a structural FP procedure based on the Tanimoto similarity coefficient were all employed in the analysis, yielding a breadth value of $\Delta = 132/159 = 0.83$. This shows that the approach covered a wider region of CS than could have been achieved using a single similarity (distance) measure. Moreover, the ratio of the fraction of actives in the three samples, $f_{\text{sample}} = 23/132 = 0.174$, to the fraction of actives obtained from a random sample of the database, $f_{\text{background}} \approx 0.04$ yields an enrichment of $EF \approx 0.174/0.04 = 4.4$. Thus, nearly four and a half times as many actives were obtained than would be expected by randomly sampling and screening compounds in the DB—more details can be obtained in the paper.

While this enhancement may not seem like a significant improvement over background, it is if a *Las Vegas model* of drug discovery is considered. As is true for many of the gambling activities in Las Vegas such as roulette and craps, the odds of winning are "shaved" slightly in the House's favor. Given that enough people place bets, statistically the House will almost certainly win over time. This has a close parallel to the HTS in drug discovery. If the odds of finding actives are even slightly better than those for random screening, and if enough compounds are screened, active compounds will almost certainly be found given that the compound DB is not highly biased, that is filled with biologically unsuitable compounds. Even an enhanced enrichment factor of two can still yield actives, but the smaller the factor the more compounds that need to be screened.

Target (Activity) Class-Specific Similarity Searching The basic idea behind target (activity) class-specific[23] similarity searching is that particular feature descriptors may exhibit some bias for specific classes of bioactivity such as, for example, HMG Co-A Reductase inhibitors, COX2 inhibitors, and 5HT (serotonin) receptor ligands. Since work in this area is based primarily on molecule-independent structural FPs, their bit positions can be unequivocally associated with specific structural features. The probability that a given feature is associated with a specific activity is estimated essentially by computing its relative frequency of occurrence in the set of molecules associated with that target class. Bits associated with features having high probabilities of occurrence, which may be called *characteristic bits*, are generally, but not always, weighted in some fashion to further emphasize their importance in subsequent similarity analyses; weighting can be accomplished in a number of ways (*vide infra*).

This approach to target class-specific similarity searching, called *reverse fingerprinting* by Williams [234], has also been carried out in a number of other laboratories [235–242]. The application of methods utilizing "nontraditional" structural fragments [234, 237, 239] have shown promise, but none of the earlier methods including these have addressed the issue of interdependencies among structural descriptors. Two papers from the Bajorath group [240, 241] that show promise have taken steps in this direction.

Based on a growing amount of data that show that *compound* and *target promiscuity* is more ubiquitous than had earlier been suspected may present significant challenges to the development of robust target class-specific similarity searching that is difficult to overcome (See Sect. 1.3.1 for further discussion).

1.4 Summary and Conclusions

Over the past two decades, computational methods have been playing an ever-increasing role in drug discovery research due especially to the burgeoning amount of data being generated by ever faster and more powerful experimental techniques. Three concepts, molecular similarity, CS, and activity/property landscapes, in some fashion underlie all of these methods—the current work addresses molecular/structural similarity and CS, two important pillars supporting the edifice of chemical informatics.

[23] In order to simplify discussion, the terminology "target class specific" will be used in the remainder of this section.

Similarity is probably one of the most ubiquitous concepts in many human endeavors. Hence, it is no surprise that it also plays a significant role in many aspects of chemical informatics. And, as is essentially true in all conscious and subconscious applications of the concept, however, what precisely it is remains somewhat a mystery since "similarity like pornography is difficult to define but you know it when you see it" [10]. The inherent subjectivity of similarity poses significant problems in chemical informatics since its application in this field is, in many cases, carried out computationally. Two key issues that then must be addressed are how to represent the relevant chemical or molecular information and how to compute an effective measure of similarity from that information. This has been covered extensively for a variety of 2-D similarity measures in Sect. 1.2 that, due primarily to their generally higher computational speeds, are by far the most popular similarity measures in use today. Surprisingly, perhaps, 2-D similarity measures perform comparably or better than many 3-D measures in a variety of cheminformatics tasks, one reason along with their higher computational speeds that accounts for their popularity.

An interesting extension of similarity-based methods that shows promise involves combining similarity values using data fusion techniques that have been applied in many engineering applications. In some cases, fused similarity values have been shown to yield significantly improved results. This is especially true of an approach called group fusion, which is based on computing the similarity of compounds in a large DB with respect to a number of reference compounds using a single similarity measure. The similarity or rank values for each DB compound are then fused to yield a single similarity score or ranking. The resulting list provides a set of compounds such that those of higher rank can be selected, for example, for follow-on screening.

A discussion presented in Sect. 1.2.4 suggests a rationale, based on the surprising prevalence of similarity cliffs, as to why group fusion appears to perform better in similarity searches than the use of a single similarity measure or the fusion of multiple similarity measures, both carried out with respect to a single reference compound. This is understandable since the relatively common occurrence of similarity cliffs, which arise when two structurally dissimilar compounds have similar activities in a given assay, suggests that active compounds may in many cases be more widely dispersed through CSs than heretofore had been suspected. Moreover, the fact that the more dissimilar the set of reference compounds the better the results of group fusion similarity searches supports this contention. An unresolved issue with this approach to similarity searching is the need for multiple active reference compounds, a situation that may not be realized in the initial phase of an HTS campaign.

Aside from its computational uses in chemical informatics, similarity also plays a significant *perceptual* role in many aspects of chemistry. This clearly is the case in medicinal chemistry where chemists address the question of "what to make next" by inferring new structures for synthesis based on the structures of active and inactive compounds considered earlier. There are, of course, many other such examples one can think of, all of which raise the issue as to whether computed similarities are comparable to those perceived by chemists.

As discussed in Sect. 1.2.5, the similarity scale, which generally is taken to lie on the unit interval [0,1] of the real line, is not uniform in terms of human perception.

Humans excel at comparing very similar objects, just as chemists excel at recognizing very similar molecules. However, at some point, as objects become less and less similar, humans can no longer discern how dissimilar they are to one another, only that they are very dissimilar. This is not entirely the case computationally since computers make no value judgments; they implement specific algorithms, although a caveat discussed in Sect. 1.2.1.4 shows that computational algorithms can also exhibit idiosyncratic behaviors such as the size-dependent behavior of FP-based similarity coefficients.

A possible reason for this disparity between chemists' perceived similarity values and those obtained computationally is seen in the expressions for Tanimoto similarity and dissimilarity given in Eqs. (1.8) and (1.21), respectively. Since the denominators in both equations are identical, it is their respective numerators that determined the difference in these two coefficients. In the case of similarity, the numerator is based on the number of features in *common* in the two molecules, while in the case of dissimilarity, the numerator is based on the number of features *unique* to each molecule. Unique features, that is, features in one molecule but not in the other, are more difficult for humans to perceive than features common to both molecules. Thus, cases of low similarity (few features in common) or high dissimilarity (more unique features) are difficult for humans to perceive. Clearly, the perceptual issue goes beyond the mathematical complementarity exhibited by Eq. (1.19). Importantly, these arguments provide a mechanism that may account for the limited correspondence between computed and perceived similarities and dissimilarities.

The notion of CS is closely related to that of similarity. Section 1.3 provides a discussion of three possible representations of CSs, namely, coordinate based (Sect. 1.3.2), cell based (Sect. 1.3.3), and graph or network based (Sect. 1.3.4). The first two are well known in the chemical informatics field. The last is not, although networks are being employed to describe a growing number of chemically related systems such as those, for example, describing protein–protein interactions, drug–target relationships, and pharmacological space. The network-based approach, which opens up new ways to investigate the nature of CSs, has two distinct advantages, namely, it is inherently discrete and it provides an intuitive representation of these spaces. Unfortunately, very few papers describing network-based representations of CSs have been published, but the power of this approach would seem to auger well for its future application in chemical informatics. In this regard, a new graph-based DB scheme that may provide a powerful approach for treating CSs, is gaining recognition in the computer field.

Each of the three CS representations has its strengths and weaknesses with regard to the types of applications for which they are best suited. A number of examples such as:

- Comparing compound DBs
- Selecting chemically diverse subsets
- Augmenting DBs through compound acquisition
- Similarity searching—2-D LBVS

are presented in Sect. 1.3.5 to illustrate this point.

The need for computational methods that can characterize relationships among sets of molecules is clearly manifest, especially in this age of massive and rapidly growing compound DBs. And although imperfect almost by their very nature, similarity-based methods provide the means for addressing this critical need. These methods also provide the means for constructing CSs that help to unify the chemical universe in an intuitive and computationally powerful way. Both notions are now beginning to be applied in fields outside of chemical informatics such as materials science and engineering laying the groundwork for future applications in food science and related fields.

Acknowledgments Many thanks to Prof. Dr. Jurgen Bajorath and his Group, especially Dr. Martin Vogt, for numerous helpful discussions. Thanks also to Dr. Jose Medina-Franco for providing Figs. 1.9 and 1.10 and to Drs. Mic Lajiness and Veer Shanmugasundaram for providing the data and for constructing Figs. 1.8 and 1.17, and lastly to Dr. Vijay Gokhale for reading and commenting on the entire manuscript.

References

1. Ruddigkeit L, van Deursen R, Blum LC, Reymond J-L (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GBD-17. J Chem Inf Model 52:2864–2875
2. Bohacek RS, McMartin C, Guida WC (1996) The art and practice of structure-based drug design: a molecular modeling perspective. Med Res Rev 16:3–50
3. Virshup AM, Contreras-Garcia J, Wipf P, Yang W, Beratan DN (2013) Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. J Amer Chem Soc 135:7296–7303
4. Wassermann AM, Wawer M, Bajorath J (2010) Activity landscape representations for structure-activity relationship analysis. J Med Chem 53:8209–8223
5. Iyer P, Wawer M, Bajorath J (2011) Comparison of two- and three-dimensional activity landscape representations for different compound sets. MedChemComm 2:113–118
6. Bajorath J (2012) Modeling activity landscapes for drug discovery. Expert Opin Drug Discov 7:463–473
7. Iyer P, Stumpfe D, Vogt M, Bajorath J, Maggiora GM (2013) Activity landscapes, information theory, and structure-activity relationships. Mol Inf 32:421–430
8. Vogt M, Iyer P, Maggiora GM, Bajorath J (2013) Conditional probabilities of activity landscape features for individual compounds. J Chem Inf Model 53:1602–1612
9. Rouvray DH (1990) The evolution of the concept of molecular similarity. In: Johnson MA, Maggiora GM (eds) Concepts and applications of molecular similarity, chapter 2. Wiley, New York
10. Medina-Franco JL, Maggiora GM (2014) Molecular similarity analysis. In: Bajorath J (ed) Chemoinformatics in drug discovery: concepts, methods, and tools for drug discovery, chapter 15. Wiley, New York
11. Mendeleev D (1869) J Russ Phys Chem Soc 1:60
12. Meyer L (1870) Ann Suppl 7:354
13. Wilkins CL, Randic M (1980) A graph theoretical approach to structure-property and structure-activity correlation. Theoret Chim Acta 58:45–68
14. Johnson M, Basak S, Maggiora G (1988) A characterization of molecular similarity methods for property prediction. Mathl Comput Model 11:630–634
15. Johnson MA, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. Wiley, New York
16. Trinajstic N (1992) Chemical graph theory, 2nd edn. CRC, Baca Raton

17. Brown RD, Martin YC (1996) Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. J Chem Inf Comput Sci 36:572–584
18. Brown RD, Martin YC (1997) The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. J Chem Inf Comput Sci 37:1–9
19. ChEMBL https://www.ebi.ac.uk/chembldb/. Accessed 1 Feb 2014
20. PubChem http://pubchem.ncbi.nlm.nih.gov. Accessed 1 Feb 2014
21. Chen J, Swamidass SJ, Dou Y, Bruand J, Baldi P (2005) ChemBD: a public database of small molecules and related chemoinformatics resources. Bioinformatics 21:4133–4139
22. DrugBank http://www.drugbank.ca. Accessed 1 Feb 2014
23. WOMBAT http://www.sunsetmolecular.com/. Accessed 1 Feb 2014
24. MDDR http://accelrys.com/products/databases/bioactivity/mddr.html. Accessed 1 Feb 2014
25. Scior JT, Bernard P, Medina-Franco JL, Maggiora GM (2007) Large compound databases for structure-activity relationships studies in drug discovery. Mini Rev Med Chem 7:851–860
26. Leach AR, Gillet VJ (2003) An introduction to chemoinformatics. Kluwer Academic, Dordrecht
27. Gasteiger J, Engel T (eds) (2003) Chemoinformatics—a textbook. Wiley-VCH, Weinheim
28. Bajorath J (ed) (2004) Chemoinformatics—concepts, methods, and tools for drug discovery. Humana, Totowa
29. Bunin BA, Siesel B, Morales G, Bajorath J (2006) Chemoinformatics: theory, practice, and products. Springer, New York
30. Bajorath J (ed) (2011) Chemoinformatics and computational chemical biology. Humana, New York
31. Willett P, Barnard JM, Downs GM (1998) Chemical similarity searching. J Chem Inf Comput Sci 38:983–986
32. Bender A, Glen RC (2004) Molecular similarity: a key technique in molecular informatics. Org Biomol Chem 2:3204–3218
33. Willett P (2009) Similarity methods in chemoinformatics. Annu Rev Inf Sci Technol 43:3–71
34. Maggiora GM, Vogt M, Stumpfe D, Bajorath J (2014) Molecular similarity in medicinal chemistry. J Med Chem 57:3186–3204
35. Lipinski C, Hopkins A (2004) Navigating chemical space for biology and medicine. Nature 432:855–861
36. Dobson CM (2004) Chemical space and biology. Nature 432:424–428
37. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldman H (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Nat Acad Sci U S A 102:17272–17277
38. Reymond J-L, van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Med Chem Comm 1:30–38
39. Reymond J-L, Awale M (2012) Exploring chemical space for drug discovery using the chemical universe database. ACS Chem Neurosci 3:649–657
40. Yu MJ (2013) Druggable chemical space and enumerative combinatorics. J Cheminformatics 5:19. doi:10.1186/1758–2964-5–19
41. Maggiora GM, Shanmugasundaram V (2011) Molecular similarity measures. In: Bajorath J (ed) Chemoinformatics and computational chemical biology, Chapter 2. Humana, New York
42. Baldi P, Benz RW, Hirschberg DS, Swamidass SJ (2007) Lossless compression of chemical FPs using integer entropy codes improves storage and retrieval. J Chem Inf Model 47:2098–2109
43. MACCS structural keys. Symyx software: San Ramon2005
44. Barnard JM, Downs GM (1997) Chemical fragment generation and clustering software. J Chem Inf Comput Sci 37:141–142
45. Carhart RE, Smith DH, Venkataraghaven R (1985) Atom pairs as molecular features in structure-activity studies. J Chem Inf Comput Sci 25:64–73
46. Rogers D, Hahn M (2010) Extended-connectivity FPs. J Chem Inf Model 50:742–754

47. Daylight IS (2014) Fingerprints—screening and similarity. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html. Accessed 2 Feb 2014
48. ChemAxon (2014) ECFP—extended connectivity fingerprints. http://www.chemaxon.com/jchem/doc/user/ECFP.html. Accessed 3 Feb 2014
49. Hu Y, Lounkine E, Bajorath J (2009) Improving the search performance of extended connectivity fingerprints through activity-oriented feature filtering and application of a bit-density-dependent similarity function. ChemMedChem 4:540–548
50. Glen RC, Bender A, Arnby CH, Carlsson L, Boyer S, Smith J (2006) Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs 9:199–204
51. Arif SM, Holiday JD, Willett P (2009) Analysis and use of fragment-occurrence data in similarity-based virtual screening. J Comput Aided Mol Des 23:6655–6668
52. Arif SM, Hert J, Holliday JD, Malim N, Willett P (2009) Enhancing the effectiveness of FP-based virtual screening: Use of turbo similarity searching and of fragment frequencies of occurrence. In: Kadirkamanathan V, Sanguinetti G, Girolami M, Niranjan M, Noirel J (eds) Pattern recognition in bioinformatics—Proceedings 4th IAPR international conference, Springer, Berlin, pp 404–414
53. Arif SM, Holiday JD, Willett P (2010) Inverse frequency weighting of fragments for similarity-based virtual screening. J Chem Inf Model 50:1340–1349
54. Willett P, Winterman V (1986) A comparison of some measures for the determination of inter-molecular structural similarity measures of inter-molecular structural similarity. Quant Struct Act Relat 5:18–25
55. Tversky A (1977) Features of similarity. Psychol Rev 84:327–352
56. Maggiora GM, Petke JD, Mestres J (2002) A general analysis of field-based molecular similarity indices. J Math Chem 31:251–270
57. Chen X, Brown F (2007) Asymmetry of chemical similarity. ChemMedChem 2:180–182
58. Wang Y, Eckert H, Bajorath J (2007) Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. ChemMedChem 2:1037–1042
59. Lipkus AH (1999) A proof of the triangle inequality for the Tanimoto distance. J Math Chem 26:263–265
60. Hankerson D, Harris GA, Johnson Jr PD (1998) Introduction to information theory and data compression. CRC, Boca Raton
61. Flower DR (1988) On the properties of bit string based measures of chemical similarity. J Chem inf Comput Sci 38:379–386
62. Lajiness M (1990) Molecular similarity–based methods for selecting compounds for screening. In: Rouvray D (ed) Computational chemical graph theory. Nova Science, pp 299–316
63. Lajiness MS (1997) Dissimilarity-based compound selection techniques. Perspect Drug Disc Design 7/8:65–84
64. Dixon SL, Koehler RT (1999) The hidden component of size in two-dimensional fragment descriptors: side effects on sampling in bioactive libraries. J Med Chem 42:2887–2900
65. Fligner MA, Verducci JS, Blower PE (2002) A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings. Technometrics 44:110–119
66. Godden WJ, Xue L, Bajorath J (2000) Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. J Chem Inf Comput Sci 40:163–166
67. Holliday JD, Salim N, Whittle M, Willett P (2003) Analysis of size dependence of chemical similarity coefficients. J Chem Inf Comput Sci 43:819–828
68. Marshall AG (1978) Biophysical chemistry. Wiley, New York
69. Hehre WJ, Radom L, Schleyer PvR, Pople JA (1986) Ab initio molecular orbital theory. Wiley, New York
70. Devillers J, Balaban AT (eds) (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach Science, New York

71. Martin Y (2010) Quantitative drug design–a critical introduction, 2nd edn. CRC, New York
72. Todeschini R, Consonni V (2009) Molecular descriptors for chemoinformatics, vol 1, 2nd edn. Wiley-VCH, Weinheim
73. Guha R, Willighagen E (2010) A survey of quantitative descriptions of molecular structure. Curr Top Med Chem 12:1946–1956
74. Labute P (2000) A widely applicable set of descriptors. J Mol Graph Model 18:464–467
75. Labute P (2004) Derivation and application of molecular descriptors based on approximate surface area. In: Bajorath J (ed) Chemoinformatics: concepts, methods, and tools for drug discovery, Chapter 8. Humana, Totowa
76. Pearlman RS, Smith KS (2002) Novel software tools for chemical diversity. 3D QSAR in drug design: three-dimensional quantitative structure-activity relationships 2:339–353
77. Pearlman RS, Smith KM (1999) Metric validation and the receptor-relevant subspace concept. J Chem Inf Comput Sci 39:28–35
78. Pearlman RS (1995) Diverse solutions user's manual. University of Texas, Austin
79. Burden F (1989) Molecular identification number for substructure searches. J Chem Inf Comput Sci 29:225–227
80. Menard PR, Mason JS, Morize I, Bauerschmidt S (1998) Chemistry space metrics in diversity analysis. J Chem Inf Comput Sci 38:1204–1213
81. Schnur D (1999) Design and diversity analysis of large combinatorial libraries using cell-based methods. J Chem Inf Comput Sci 39:36–45
82. Mason JS, Beno BR (2000) Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure-based diversity. J Mol Graphics Model 18:438–451
83. Stanton DT (1999) Evaluation and use of BCUT descriptors in QSAR and QSPR studies. J Chem Inf Comput Sci 39:11–20
84. Pirard B, Pickett SD (2000) Classification of kinase inhibitors using BCUT descriptors. J Chem Inf Comput Sci 40:1431–1440
85. González MP, Terán C, Besada TM, González-Moa MJ (2005) BCUT descriptors to predicting affinity toward A3 adenosine receptors. Bioorg Med Chem Lett 15:3491–3495
86. Shanmugasundaram V, Maggiora GM, Lajiness MS (2005) Hit-directed nearest neighbor searching. J Med Chem 48:240–248
87. Hodgkin EE, Richards WG (1987) Molecular similarity based on electrostatic potential and electric field. Int J Quantum Chem Quantum boil Symp 14:105–110
88. Sheridan RP, Kearsely SK (2002) Why do we need so many chemical similarity search methods? Drug Discov Today 7:903–911
89. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physicochemical property descriptors. J Chem Inf Comput Sci 36:11–127
90. Sheridan RP, Miller MD, Underwood DJ, Kearsley SK (1996) Chemical similarity using geometric atom pair descriptors. J Chem Inf Comput Sci 36:128–136
91. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Comparison of FP-based for virtual screening using multiple bioactive structures. J Chem Inf Comput Sci 44:1177–1185
92. Whittle M, Gillet VJ, Willett P, Alex A, 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
93. Willett P (2006) Enhancing the effectiveness of ligand-based virtual screening using data fusion. QSAR Combin Sci 25:1143–1152
94. Willett P (2013) Combination of similarity rankings using data fusion. J Chem Inf Model 53:1–10
95. Joshi R, Sanderson AC (1999) Multisensor fusion: a minimal representation framework. World Scientific, Singapore
96. Hall DL, McMullen SAH (2004) Mathematical techniques in multisensory data fusion. Artech House, Boston

97. Beliakov G, Pradera A, Tomasa C (2010) Aggregation functions: a guide for practitioners. Springer, Berlin
98. Harmonic mean (2014) Wikipedia. http://en.wikipedia.org/wiki/Harmonic_mean. Accessed 7 Jan 2014
99. Cormack GV, Clark CLA, Buettcher S (2009) Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, Boston, 19–23 July 2009, pp 758–759
100. Chen B, Meuller C, Willett P (2010) Combination rules for group fusion in similarity based virtual screening. Mol Inf 29:533–541
101. Critchlow DE (1980) Metric methods for analyzing partially ranked data. Springer, New York
102. Nasr RJ, Swamidass SJ, Baldi PF (2009) Large scale study of multiple molecule queries. J Cheminform 1:7. http://www.jcheminf.com/content/1/1/7. Accessed 7 Jan 2014. doi:10.1186/1758-2946-1-7
103. Stumpf D, Bajorath J (2011) Similarity searching. WIRES Comput Mol Sci 1:260–282
104. Willett P (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discov Today 11:1046–1053
105. Gardiner EJ, Gillet VJ, Haranczyk M, Hert J, Holliday JD, Malim N, Patel Y, Willett P (2009) Turbo similarity searching: effect of FP and dataset on virtual-screening performance. Stat Anal Data Mining 2:103–114
106. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) New methods for ligand-based virtual screening:use of data-fusion and machine-learning techniques to enhance the effectiveness of similarity searching. J Chem Inf Model 46:462–470
107. Miyamoto S (1990) Fuzzy sets in information retrieval and cluster analysis. Kluwer Academic, Dordrecht
108. Edgar SJ, Holliday JD, Willett P (2000) Effectiveness of retrieval in similarity searches of chemical databases: a review of performance measures. J Mol Graph Model 18:343–357
109. Willett P (2004) Evaluation of molecular similarity and molecular diversity methods using biological data. In: Bajorath J (ed) Chemoinformatics-Concepts, methods and tools for drug discovery, Chapter 2. Humana, Towata
110. Truchon J-F, Bayly CI (2007) Evaluating virtual screening: good and bad metrics for the "early recognition" problem. J Chem Inf Model 47:488–508
111. Maggiora GM (2006) On outliers and activity cliffs—why QSAR often disappoints (Editorial). J Chem Inf Model 46:1535
112. Guha R, Van Drie J (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48:646–658
113. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry. J Med Chem 55:2932–2942
114. Stahura FL, Bajorath J (2002) Bio- and chemo-informatics beyond data management: crucial challenges and future opportunities. Drug Discov Today 7:S41–S47
115. Hu Y, Maggiora GM, Bajorath J (2013) Activity cliffs in PubChem confirmatory bioassays taking inactive compounds into account. J Comput Aided Mol Des 27:115–124
116. Lajiness MS, Maggiora GM, Shanmugasundaram V (2004) An assessment of the consistency of medicinal chemists in reviewing compound lists. J Med Chem 47:4891–4896
117. Takaoka Y, Endo Y, Yamanobe S, Kakinuma H, Okubo T, Shimazaki Y, Ota T, Sumiya S, Yoshikawa K (2003) Development of a method for evaluating drug-likeness and ease of synthesis using a data set in which compounds are assigned scores based on chemists' intuition. J Chem Inf Comput Sci 43(4)1269–1275
118. Kutchukian PS, Vasilyeva NY, Xu J, Lindvall MK, Dillon MP, Glick M, Coley JD, Brooijmans N (2012) Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. PLoS ONE 7:e48476
119. Hawkins DM, Young SS, Rusinko A III (1997) Analysis of a large structure-activity data set using recursive partitioning. Mol Inf 16:296–302

120. Chen X, Rusinko A III, Young S (1998) Recursive partitioning analysis of a large scale structure-activity data set using three-dimensional descriptors. J Chem Inf Comput Sci 38:1054–1062
121. Rusinko A III, Farmen MW, Lambert CG, Brown PL, Young SS (1999) Analysis of a large structure/biological activity data set using recursive partitioning. J Chem Inf Comput Sci 39:1017–1026
122. Wasserman S, Faust K (1997) Social network analysis. Cambridge University , New York
123. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nature Biotech 24:805–815
124. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690
125. Kesier MJ, Roth BL, Armruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25:197–206
126. Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug-target network. Nat Biotechnol 25:1119–1126
127. Watts DJ (2003) Six Degrees—the science of a connected age. WW Norton, New York
128. Barbási A-L (2003) Linked: how everything is connected to everything else, and what it means for business, science, and everyday life. Penguin, New York
129. Newman MEJ (2010) Networks an introduction. Oxford University, New York
130. Robinson I, Webber J, Eifrém E (2013) Graph databases. O'Reilly Media, Sebastopol, CA 95472
131. Peltason L, Bajorath J (2007) SAR Index: quantifying the nature of structure-activity relationships. J Med Chem 50:5571–5578
132. Namasivayam V, Iyer P, Bajorath J (2012) Exploring SAR continuity in the vicinity of activity cliffs. Chem Biol Drug Des 79:22–29
133. Hu Y, Bajorath J (2014) Exploring compound promiscuity patterns and multi-target activity spaces. Comput Struct Biotech J 9:1003–1012. http://dx.doi.org/10.5936/csbj.201401003. Accessed 23 Feb 2014
134. Medina-Franco JL (2013) Activity cliffs: facts or artifacts? Chem Biol Drug Des 81:553–556
135. Hu Y, Bajorath J (2010) Molecular scaffolds with high propensity to form multi-target activity cliffs. J Chem Inf Model 50:500–510
136. Wassermann AM, Bajorath J (2010) Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. J Chem Inf Model 50:1248–1256
137. Martin YC, Kofron JL, Traphagen LM (2002) Do structurally similar molecules have similar biological activities? J Med Chem 45:4350–4358
138. Thor and Merlin; Version 4.62; Daylight Chemical Information Systems, Inc., Irvine, CA. http://www.daylight.com. Accessed 12 Jan 2014
139. Brown RD, Martin YC (1998) An evaluation of structural descriptors and clustering methods for use in diversity selection. SAR QSAR Environ Res 8:23–39
140. Patterson DE, Cramer RD, Ferguson AM, Clark RD, Weinberger LE (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. J Med Chem 39:3049–3059
141. Steffen A, Kogej T, Tyrchan C, Engkvist O (2009) Comparison of molecular FP methods on the basis of biological profile data. J Chem Inf Model 49:338–347
142. Wikipedia. Curse of dimensionality. http://en.wikipedia.org/wiki/Curseof_dimensionality. Accessed 19 Jan 2014
143. Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading
144. Rupp M, Proschak E, Schneider G (2007) Kernel approach to molecular similarity based on iterative graph similarity. J Chem Inf Model 47:2280–2286
145. Joliffe IT (2002) Principle component analysis, 2nd edn. Springer, New York
146. Borg I, Groenen P (1997) Modern multi-dimensional scaling. Springer, New York
147. Domine D, Devillers J, Chastrette M, Karcher W (1993) Non-linear mapping for structure-activity and structure-property modeling. J Chemometr 7:227–242

148. Malinowski ER (1991) Factor analysis in chemistry, 2nd edn. Wiley, New York
149. Raghavendra AS, Maggiora GM (2007) Molecular basis sets—a general similarity-based approach for representing CSs. J Chem Inf Model 47:1328–1340
150. Kruskal J (1977) The relationship between multidimensional scaling and clustering. In: Van Ryzin J (ed) Classification and clustering. Academic, New York, pp 17–44
151. Diamantaras KI, Kung SY (1996) Principal component neural networks: theory and applications. Wiley, New York
152. Molecular Operating Environment (MOE). Chemical computing group, Montreal, Quebec, Canada. http://www.chemcomp.com. Accessed 26 Feb 2014
153. Mason JS, Good AC, Martin EJ (2001) 3-D pharmacophores in drug discovery. Curr Pharm Des 7:567–597
154. Agrafiotis DK, Xu H (2003) A geodesic framework for analyzing molecular similarities. J Chem Inf Model 43:475–484
155. Agrafiotis DK, Xu H (2002) A self-organizing principle for learning non-linear manifolds. Proc Nat Acad Sci U S A 99:15869–15872
156. Agrafiotis DK (2003) Stochastic proximity embedding. J Comput Chem 24:1215–1221
157. Xue L, Stahura FL, Bajorath J (2004) Cell-based partitioning. In: Chemoinformatics: concepts, methods, and tools for drug discovery, Chapter 9. Humana , Totowa
158. Wickens TD (2009) Multiway contingency tables analysis for the social sciences. Psychology, New York
159. Bayley MJ, Willett P (1999) Binning schemes for partition-based compound selection. J Mol Graphics Model 17:10–18
160. Rush JA (1999) Cell-based methods for sampling in high-dimensional spaces. In: Truhlar DG, Howe WJ, Hopfinger AJ, Blaney J, Dammkoehler RA (eds) Rational drug design. Springer, New York, pp 73–79
161. Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs
162. Maggiora GM, Shanmugasundaram V, Lajiness MS, Doman TN, Schultz MW (2004) A practical strategy for directed compound acquisition. In: Oprea TI (ed) Chemoinformatics in drug discovery. Wiley-VCH, Weinheim
163. Hassan M, Bielawski JP, Hempel JC, Waldman M (1996) Optimization and visualization of molecular diversity of combinatorial libraries. Mol Divers 2:64–74
164. Sotriffer C, Manhold R, Kubinyi H, Folkers G (2011) Virutal screening—principles, challenges, and practical guidelines. Wiley, New York
165. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM (2004) Protein interaction networks from yeast to human. Curr Opin Struct Biol 14:292–299
166. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target networks from the integration of chemical and genomic spaces. Bioinformatics 24:1232–1240
167. Zhao S, Li S (2010) Network-based relating pharmacological and genomic spaces for drug target identification. PLoS ONE 5(7):e11764. doi:10.1371/journal.pone.0011764
168. Tanaka N, Ohno K, Niimi T, Moritomo A, Mori K, Orita M (2009) Small-world phenomena in chemical library networks: application to fragment-based drug discovery. J Chem Inf Model 49:2677–2686
169. Krein MP, Sukumar N (2011) Exploration of the topology of chemical spaces with network measures. J Phys Chem A 115:12905–12918
170. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. J Med Chem 51:6075–6084
171. Ripphausen P, Nisius B, Wawer M, Bajorath J (2011) Rationalizing the role of SAR tolerance for ligand-based virtual screening. J Chem Inf Model 51:837–842
172. Stumpfe D, Dimova D, Bajorath J (2014) Composition and topology of chemical spaces with network measures. J Chem Inf Model 54:451–461
173. Benz RW, Swamidass SJ, Baldi P (2008) Discovery of power-laws in chemical space. J Chem Inf Model 48:1138–1151

174. Oprea TI, Gottfries J (2001) Chemography: the art of navigating in chemical space. J Comb Chem 3:157–166
175. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
176. Harary F (1969) Graph theory. Addison-Wesley, Reading
177. Bolla M (2013) Spectral clustering and biclustering—learning large graphs and contingency tables. Wiley, New York
178. Kolaczyk ED (2009) Statistical analysis of network data—methods and models. Springer, New York
179. Liu B (2011) Web data mining: exploring hyperlinks, contents, and usage data. Springer, Heidelberg
180. van Steen M (2010) Graph theory and complex networks—an introduction. Maarten van Steen
181. Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. Proc Nat Acad Sci U S A 97:11149–11152
182. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512
183. Devore JL, Berk KN (2011) Modern mathematical statistics with applications. Springer, New York
184. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compoundsfor virtual screening. J Chem Inf Model 45:177–182
185. Rajan K (ed) (2013) Informatics for materials science and engineering: data-driven discovery for accelerated experimentation and applications. Elsevier, New York
186. Hudson BD, Hyde RM, Rahr E, Wood J, Osman J (1996) Parameter based methods for compound selection from chemical databases. Quant Struct-Act Relat 15:285–289
187. Holliday JD, Willett P (1996) Definitions of "dissimilarity" for dissimilarity-based compound selection. J Biomolec Screen 1:145–151
188. Menard PR, Lewis RA, Mason JS (1998) Rational screening set design and compound selection: cascaded clustering. J Chem Inf Comput Sci 38:497–505
189. Young SS, Lam RLH, Welch WJ (2002) Initial compound selection for sequential screening. Curr Opin Drug Discov Devel 5:422–427
190. Waldman M, Li H, Hassan M (2000) Novel algorithms for the optimization of molecular diversity of combinatorial libraries. J Mol Graph Model 18:412–426
191. Agrafiotis DK (1998) Diversity in chemical libraries. In Schleyer PvR, Allinger NL, Clark T, Gasteiger J, Kollman PA, Schaefer HF III, Schreiner PR (eds) The Encyclopedia of Computational Chemistry, pp 742–761, John Wiley & Sons, Chichester
192. Shanmugasundaram V, Maggiora G (2011) Application of Shannon-like diversity measures to cell-based chemistry spaces. J Math Chem 49:342–355
193. Willett P (2000) Chemoinformatics—similarity and diversity in chemical libraries. Curr Opin Biotechnol 11:85–88
194. Willett P (2004) Evaluation of molecular similarity and molecular diversity methods using biological activity data. In: Bajorath J (ed) Chemoinformatics: concepts, methods, and tools for drug discovery, Chapter 2. Springer, New York
195. Martin Y (ed) (2001) Diverse viewpoints on computational aspects of molecular diversity. J Comb Chem 3:231–250
196. Matter H (1997) Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors. J Med Chem 40:1219–1229
197. Dunbar JB (2000) Compound acquisition strategies. Pac Symp Biocomput 5:552–562
198. Olah MM, Bologa CG, Oprea TI (2004) Strategies for compound selection. Curr Drug Discov Technol 1:211–220
199. Ma C, Lazo JS, Xie X-Q (2011) Compound acquisition and prioritization algorithm for constructing structurally diverse compound libraries. ACS Comb Sci 13:223–231

200. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimates solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 46:3–26
201. Petit J, Meurice N, Kaiser C, Maggiora G (2012) Softening the rule of five—where to draw the line? Bioorg Med Chem 20:5343–5351
202. Bickerton GR, Pailini GV, Besnard J, Muresan S, Hopkins AL (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98
203. Klebe G (ed) (2000) Virtual screening: an alternative or complement to high throughput screening? Kluwer Academic, Dordrecht
204. Varnek A, Tropsha A (eds) (2008) Chemoinformatics approaches to virtual screening. RSC Publishing, Cambridge
205. Böhm H-J, Schneider G, Kubinyi H, Manhold R, Timmerman H (eds) (2008) Virtual screening for bioactive molecules. Wiley, New York
206. Bajorath J (2002) Integration of virtual and high-throughput screening. Nat Rev Drug Discov 1:882–894
207. Glen RC, Adams SE (2006) Similarity metrics and descriptor spaces—which combinations to choose? QSAR Combin Sci 25:1133–1142
208. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 12:225–233
209. Rester U (2008) From virtual reality—virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. Curr Opin Drug Discov Devel 11:559–568
210. Bajorath J (2009) Methods for ligand-based virtual screening. Frontiers Med Chem 4:1–22
211. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9:273–276
212. Geppert H, Vogt M, Bajorath J (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model 50:205–216
213. Stumpfe D, Bajorath J (2011) Similarity searching. WIREs Comput Mol Sci 1:260–282
214. Scior T, Bender A, Tresadern G, Medina-Franco JL, Mayorga KM, Langer T, Cuanalo-Contreras K, Agrafiotis DK (2012) Recognizing pitfalls in virtual screening: a critical review. J Chem Inf Model 52:867–881
215. Lavecchia A, Di Giovanni C (2013) Virtual screening strategies in drug discovery: a critical review. Curr Med Chem 20:2839–2860
216. Parker CN, Bajorath J (2006) Towards unified compound screening strategies: a critical evaluation of error sources in experimental and virtual high-throughput screening. QSAR Combin Sci 25:1153–1161
217. Yuriev E, Agostino M, Ramsland PA (2010) Challenges and advances in computational docking: 2009 in review. J Mol Recognit 24:149–164
218. Huang S-Y, Zou X (2010) Advances and challenges in protein-ligand docking. Int J Mol Sci 11:3016–3034
219. Waszkowycz B, Clark DE, Gancia E (2011) Outstanding challenges in protein-ligand docking and structure-based virtual screening. WIREs Comput Mol Sci 1:229–259
220. Mestres J, Rohrer DC, Maggiora GM (1997) A molecular field-based similarity approach to pharmacophoric pattern recognition. J Mol Graphics Model 15:114–121
221. Putta S, Lemmen l, Beroza P, Greene J (2002) A novel shape-feature based approach to virtual library screening. J Chem Inf Comput Sci 42:1230–1240
222. Koes DR, Camacho CJ (2011) Pharmer: efficient and exact pharmacophore search. J Chem Inf Model 51:1307–1314
223. Langer T (2010) Pharmacophores in drug research. Mol Inf 29:470–475
224. Mestres J, Rohrer DC, Maggiora GM (1997) MIMIC: a molecular-field matching program: exploiting applicability of molecular similarity approaches. J Comp Chem 18:934–954
225. Ballester PJ, Richards WG (2007) Ultrafast shape recognition for similarity search in molecular databases. Proc Roy Soc A 463:1307–1321

226. Hawkins P, Skillman A, Nicholls A (2007) A comparison of shape-matching and docking as virtual screening tools. J Med Chem 50:74–82
227. McGaughey GB, Sheridan RP, Baylly CI et al (2007) Comparison of topological shape and docking methods in virtual screening. J Chem Inf Model 47:1504–1519
228. Ebalunode JO, Zheng W (2009) Unconventional 2D shape similarity method affords comparable enrichment as a 3D shape method in virtual screening experiments. J Chem Inf Model 49:1313–1320
229. Yongye AB, Bender A, Martinez-Mayorga (2010) Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble. J Comput Aided Mol Des 24:675–686
230. Stanton DT, Morris TW, Siddhartha R, Parker C (1999) Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. J Chem Inf Comput Sci 39:21–27
231. Muchmore SW, Debe DA, Metz JT, Brown SP, Martin YC, Hajduk PJ (2008) Application of belief theory to similarity data fusion for use in analog searching and lead hopping. J Chem Inf Model 48:941–948
232. Swann SL, Brown SP, Muchmore SW, Patel H, Merta P, Locklear J, Hajduk PJ (2011) A unified, probabilistic framework for structure- and ligand-based virtual screening. J Med Chem 54:1223–1232
233. Sharma R, Lawrenson AS, Fisher NE et al (2012) Compound selection methods for a high-throughput screening program against a novel malarial target, PfNDH2: increasing hit rate via virtual screening methods. J Med Chem 55:3144–3154
234. Williams C (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. Mol Divers 10:311–332
235. Xue L, Stahura FL, Godden JW, Bajorath J (2001) Fingerprint scaling increases the probability if identifying molecules with similar activity in virtual screening callculations. J Chem Inf Comput Sci 41:746–753
236. Xue L, Godden JW, Stahura FL, Bajorath J (2003) Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. J Chem Inf Comput Sci 43:1218–1225
237. Schuffenhauer A, Floersheim P, Acklin P, Jacoby E (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. J Chem Inf Comput Sci 43:391–405
238. Kogej T, Engkvist Blomberg N, Muresan S (2006) Multifingerprint based similarity searches for targeted class compound selection. J Chem Inf Model 46:1201–1213
239. Batista J, Bajorath J (2008) Distribution of randomly generated activity class characteristic substructures in diverse active and database molecules. Mol Divers 12:77–83
240. Lounkine E, Auer J, Bajorath J (2008) Formal concept analysis for the identification of molecular fragment combinations specific for active and highly potent compounds. J Med Chem 51:5342–5348
241. Lounkine E, Hu Y, Batista J, Bajorath J (2009) Relevance of feature combinations for similarity searching using general or activity class-directed molecular fingerprints. J Chem Inf Model 49:561–570
242. Wassermann AM, Nisius B, Vogt M, Bajorath J (2010) Identification of descriptors capturing compound class-specific features by mutual information analysis. J Chem Inf Model 50:1935–1940