Chem Biol Drug Des 2007; **70:** 393–412

Research Article

doi: 10.1111/j.1747-0285.2007.00579.x

A Similarity-based Data-fusion Approach to the Visual Characterization and Comparison of Compound Databases

José L. Medina-Franco[1,†], Gerald M. Maggiora[1,*], Marc A. Giulianotti[2,†], Clemencia Pinilla[2] and Richard A. Houghten[2]

[1]College of Pharmacy & BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA
[2]Torrey Pines Institute for Molecular Studies, 3550 General Atomics Court, San Diego, CA 92121, USA
**Corresponding author: Gerald M. Maggiora, maggiora@pharmacy.arizona.edu*
[†]Current address: Torrey Pines Institute for Molecular Studies – PSL, 5775 Old Dixie Highway, Fort Pierce, FL 34946, USA.

A low-dimensional method, based on the use of multiple fusion-based similarity measures, is described for graphically depicting and characterizing relationships among molecules in compound databases. The measures are used to construct multi-fusion similarity maps that characterize the relationship of a set of 'test' molecules to a set of 'reference' molecules. The reference set is very general and can be made of molecules from, for example, the set of test molecules itself (the self-referencing case), from a small library or large compound collection, or from actives in a given assay or group of assays. The test set is any collection of compounds to be analyzed with respect to the specified reference set. Multiple fusion similarity measures tend to provide more information than single fusion-based measures, including information on the nature of the chemical-space neighborhoods surrounding reference-set molecules. A general discussion is presented on how to interpret multi-fusion similarity maps, and several examples are given that illustrate how these maps can be used to compare compound libraries or collections, to select compounds for screening or acquisition, and to identify new active molecules using ligand-based virtual screening.

Key words: combinatorial libraries, compound acquisition, compound selection, data visualization, diversity analysis, fusion-based similarity, ligand-based virtual screening, multi-fusion similarity maps

Received 16 August 2007, revised and accepted for publication 11 September 2007

Rapid growth in the size and availability of compound databases (1) has created the need for effective computational tools with which to analyze them. Many of these tools are based on the concept of chemical space (2–4), which provides a suitable framework for characterizing and comparing databases. As in the case of abstract mathematical spaces (5), a chemical space is made up of a set of points, representing the molecules (objects) in the space, and one or more relations (e.g. distances or similarities and associated properties such as bioactivities) on the set of points. Chemical spaces are typically represented as co-ordinate-based spaces, where each co-ordinate axis is defined by some type of molecular descriptor (6), but this is not absolutely required. For example, pairwise similarities (7) among sets of molecules are also suitable, and such 'coordinate-free' spaces can be transformed into co-ordinate-based spaces by a variety of methods (8–12). Unfortunately, chemical spaces tend to be of relatively high-dimension so as to preclude their visual depiction in two or three dimensions, but this can overcome to some extent by a number of available methods, affording the possibility that visual data analysis, a useful tool in cheminformatics, can be carried out.

Although visual data analysis is quite useful, in many cases certain features in a given chemical space can become obscured. For example, the relative ordering of nearest neighbors (NNs; e.g. first-NN, second-NN, third-NN, etc.) with respect to a given query molecule or the magnitude of similarity between two molecules may be difficult to discern visually. In addition, it is not always possible to represent all of the features of high-dimensional chemical spaces in lower-dimensional spaces. Thus, a goal of the current work was the development of new, low-dimensional representations that capture features of high-dimensional chemical spaces that may be missed in lower-dimensional chemical-space representations. Such novel low-dimensional representations will clarify these relationships and, thus, enhance scientific visualization.

A number of questions arise in the analysis of compound databases in chemical space: How diverse are the compounds in the database? What is the distribution of compounds within chemical space? Where are the densely and sparsely populated regions and what molecules are found there? Where are the active compounds located? How similar are the compounds in one database to those in another? Which compounds should be selected for a screening campaign? Which compounds should be purchased to augment a compound collection?

There are a variety of methods for tackling these questions. Two approaches, cell-based and distance- or similarity-based, are commonly used to address these questions. The first approach, which is generally applied in low-dimensional chemical spaces, involves partitioning the space into a set of multi-dimensional hypercubes, typically of equal size (13–16), although cells with more complex geometries have also been discussed (17,18). Numerous applications of cell-based methods have demonstrated their usefulness in many aspects of chemical-space analysis (19). The second approach includes a wide variety of methods that make use of information on the distance or similarity between pairs of compounds within a chemical space. The most important of these are clustering (20–24) and NN methods (25–30). Other methods based information theory (31) and self-organizing maps (32) have also been applied.

The current work adopts a NN approach and uses molecular similarity to characterize the degree to which a given molecule is a NN (e.g. first-NN, second-NN, third-NN, etc.) of a specific reference molecule or set of reference molecules. As chemical spaces are highly representation dependent, two different representations of the same set of molecules can give rise to entirely different chemical spaces (7). Thus, the results obtained with respect to one representation will not, in general, be in accord with those obtained with another representation. This general but daunting feature of chemical spaces was clearly pointed several years ago in the work of Sheridan and Kearsely (33). A variety of approaches have been considered for ameliorating the representation problem, such as by combining information from multiple chemistry spaces (30) or by using multiple similarity measures (34).

Over the last few years, data-fusion methods have been popularized by the Willett group at the University of Sheffield (35–37). These methods were developed in the engineering field as a way to combine data from a number of sources (38,39). In chemistry, data fusion, called *similarity fusion*, is typically used as a way to identify NNs of a single reference molecule by combining the rank ordering or similarities of NNs obtained from multiple similarity measures with respect to the same reference molecule. Numerous combining rules have been developed to accomplish this task. The most common ones are max-fusion and sum-fusion. In max-fusion, the molecule with the highest ranking or largest similarity score with respect to a given reference molecule, taken over all of the similarity measures, is chosen. In sum-fusion, the molecule with highest score obtained by summing the rankings or similarities with respect to a given reference molecule over all of the similarity measures, is chosen. Mean-fusion is simply sum-fusion normalized by the number of similarity measures used. Quantile-based fusion rules offer an alternative to the combining rules described above. Preliminary studies in our laboratories show that mean-fusion performs in an analogous manner to median or other quantile-based fusion measures (e.g. 90-th percentile), and thus, these measures are not considered further in this work.

The Willett group has also developed a modified approach called *group fusion* that applies the data-fusion procedure to multiple reference or query molecules (35,37), which is based on several earlier works (40,41). In contrast to similarity fusion, only a single similarity method is used in group fusion, but the sum, average, etc. similarity is computed with respect to the entire set of reference molecules. As shown by Hert *et al.* (35) in their seminal work, the use of similarity values generally produced results that were superior to those obtained based on rankings. Thus, in this work only similarity scores are considered. A recent extension of this method, called *turbo-similarity searching*, assumes that all NNs of known actives are also active, and then applies group fusion to this augmented reference set (25,42).

The different implementations of fusion-based similarity are designed to enhance the effectiveness of ligand-based virtual screening over that obtained by more 'traditional' similarity methods, although this may not always be the case. Similarity-based approaches, in general, and fusion-based approaches, in particular, are usually evaluated by 'numerical experiments' that attempt to assess the recall rate or some related measure such as the area under a ROC curve (43), which measures how effective a particular procedure is in recovering the remaining actives from a data set of known actives given a subset of these actives as reference molecules. Several studies are based on the evaluation of multiple-fusion rules to identify the 'best rule' (35,44,45).

In this study, fusion-based similarity is applied to the visual characterization and comparison of compound databases by employing multiple fusion rules in two-dimensional maps called multi-fusion similarity (MFS) maps. The focus of the approach is on the relationship of molecules in a 'test' set to molecules in a given reference set, which need not be biologically active, in contrast to the situation in typical virtual-screening applications. The fusion data generated in a given study is typically plotted in two dimensions, where the ordinate represents the max-fusion values and the abscissa the average-fusion values. Each of the points in the plot is associated with a specific molecule in the test set, and its position is determined by the corresponding fusion values computed with respect to molecules in the reference set.

The combination of fusion rules presented in this paper expands the current applications of fusion-based similarity. Combining fusion rules in low-dimensional maps provides powerful visual tools not only in ligand-based virtual screening applications, but also in diversity analysis of compound libraries or collections; comparing compound libraries or collections; compound acquisition; and design of focused and diverse combinatorial libraries as summarized in Table 1.

The test and the reference sets can be obtained from small libraries, combinatorial libraries, large compound collections, sets of active compounds, etc., or any combination of them. In the present study, a number of compound collections/libraries covering a range of sizes and diversities were investigated. These included a collection of known drug molecules ('DrugBank'; 46), a diverse collection of molecules available from the National Cancer Institute ('NCIDiv'[a]) a library containing molecules active in a number of CNS assays ('CNS'[b]) and a library of molecules active in a kappa opioid receptor assay ('Kappa') obtained from the World of Molecular Bioactivity (WOMBAT; 47). The characteristics of the four data sets are summarized in Table 2; further discussion of the properties is provided in Section Overview of compound databases used in this work.

Table 1: Potential applications of multi-fusion similarity maps

Applications	Reference set	Section discussed
Combinatorial library design	Enumerated combinatorial library and/or set of actives	General overview
Comparing compound collections	Compound collections	Comparing compound collections
Diversity analysis (profile)	Self-reference	Comparing compound collections
Compound selection	Existing compound collection	Compound selection and acquisition
Compound acquisition	Existing compound collection	Compound selection and acquisition
Ligand-based virtual screening	Actives	Ligand-based virtual screening

Table 2: Databases employed in this study

			Similarity*		
Database	General contents	Size	Mean	SD	Reference
DrugBank	Database with a wide range of drugs including FDA-approved and experimental drugs	1055	0.314	0.130	46
NCI diversity	Diverse collection of molecules available from the National Cancer Institute	1990	0.282	0.117	[a]
CNS	Collection of FDA-approved psychiatric drugs	77	0.361	0.159	[b]
Kappa	Library of molecules active in a kappa opioid receptor assay obtained from WOMBAT	196	0.604	0.139	47

*Tanimoto similarity with MACCS key fingerprints were used to compute similarity (see Section Overview of compound databases used in this work for further discussion).

A number of analyses are carried out that illustrate the multi-fusion approach. These include intra-library or intra-collection diversity analysis, where both the test and reference sets are taken to be identical (the so-called 'self-referencing' case), inter-library or inter-collection comparisons, selection of compounds for screening and acquisition, ligand-based virtual screening. In all of the examples, emphasis is placed on the graphical depiction of the multiple, fusion-based similarity data in ways that facilitate visual data analysis. Section Methodology presents a description of MFS maps, the method to obtain these maps, and general guidelines for their interpretation. Section Overview of compound databases used in this work provides an overview of the compound databases involved in this study. Section Results and Discussion describes a number of applications of the approach (vide supra). Section Conclusions concludes with a summary of the work and several conclusions.

Methodology

Basis of the multi-fusion approach: generating multi-fusion similarity maps

In all computations carried out in this work, molecules are represented by 2D MACCS key fingerprints[c] as implemented in the Molecular Operating Environment (MOE) program[d] and the similarity of the *i*-th and *j*-th molecules is computed using the well-known Tanimoto similarity coefficient (7,48),

$$T(i,j) = \frac{c}{a+b-c} \qquad (1)$$

where *a* and *b* are the number of fragment bits corresponding to the *i*-th and *j*-th molecules and *c* is the number of fragment bits common to both molecules. Despite some caveats related to size-dependent effects (49,50), the Tanimoto coefficient is the measure of choice to asses the molecular similarity of molecules based on 2D fingerprints, because on its extensive usage in a wide variety of studies (37).

Fusion similarity scores ('fused scores') are calculated for each of the *t* molecules in the test set, $i = 1, 2, ..., t$, with respect to all *n* molecules in the reference set, $r = 1, 2, ..., n$, using a particular fusion rule (35,37)

max-fusion:

$$F_n^{\max}(i) = \max_{r=1}^{n}\{T(i,r)\} \;\; ; \;\; i = 1, 2, ..., t \qquad (2)$$

sum-fusion:

$$F_n^{\text{sum}}(i) = \sum_{r=1}^{n} T(i,r) \;\; ; \;\; i = 1, 2, ..., t \qquad (3)$$

mean-fusion:

$$F_n^{\text{mean}}(i) = \frac{1}{n}\sum_{r=1}^{n} T(i,r) = \frac{1}{n}F^{\text{sum}}(i) \quad i = 1, 2, \ldots, t \qquad (4)$$

Mean-fusion corrects sum-fusion scores for the size of the reference set, and thus is appropriate when multiple reference sets of significantly different sizes are being considered, as is the case here. In addition, the inequality given in eqn 5, which can be derived from eqns 2 and 4,

$$F_n^{\text{mean}}(i) \;\le\; F_n^{\max}(i) \;\; ; \;\; i = 1, 2, ..., t \qquad (5)$$

shows that mean-fusion similarity may be equal to but is never greater than the corresponding max-fusion value. Equality occurs in the case of 'classical' similarity searching, where the reference set consists of a single compound. In such a case, all of the points lie on the diagonal in Figure 1, and both fusion-based similarity measures yield identical results.

As illustrated in Figure 1, MFS maps are two-dimensional. The ordinate corresponds to the max-fusion similarity rule and the abscissa to either the sum- or mean-fusion similarity rule. The latter is chosen in this work as it is independent of the size of the reference set (vide supra). Each point on the plot represents a test set molecule, whose location depends on the values of the fusion-based similarity measures defined in eqns 2–4. The diagonal line in Figure 1 divides the figure into two, equal-size triangular regions. Because of the relationship given in eqn 5, test set molecules cannot lie within the lower triangular region shaded in grey, as it is not possible for test set molecules to have both low max-fusion and high mean-fusion similarity values. Points that lie in the upper triangular region, including the diagonal line, satisfy eqn 5 and correspond to allowable pairs of fusion similarity values. The triangular shape of the allowed region in Figure 1 shows that as max-fusion similarity decreases, the corresponding range of allowed mean-fusion values becomes smaller. In this case, little is gained using the multi-fusion approach. As max-fusion similarity increases toward unity, the corresponding range of allowed mean-fusion values becomes quite large indicating, as will be discussed in Section Interpreting multi-fusion similarity maps, the potential for a significant variation in the geometric relationship of test set molecules to molecules in the reference set. When a subset of test-set molecules with high max-fusion similarities also possess a range of mean-fusion values, the power of the multi-fusion approach is greatest as the discriminating power of mean-fusion similarity becomes significant. On the other hand, when the range of mean-fusion values is limited, the discriminating power of the mean-fusion similarity also is diminished. Four prototypical examples described in Section Interpreting multi-fusion similarity maps illustrate these points in more detail, including the application of MFS maps in the self-reference case.

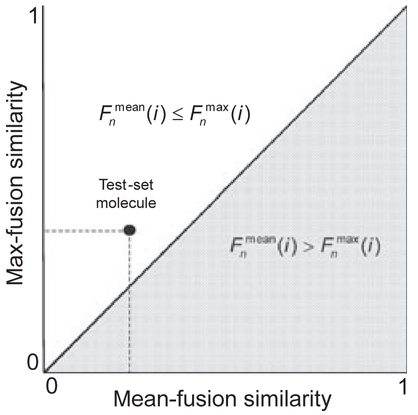

1
Max-fusion similarity
$F_n^{\text{mean}}(i) \le F_n^{\text{max}}(i)$
Test-set molecule
$F_n^{\text{mean}}(i) > F_n^{\text{max}}(i)$
0
0
1
Mean-fusion similarity

Figure 1: A schematic depiction of the general form of an MFS map. All of the test-set molecules lie in the upper triangular portion of the map and satisfy the relationship $F_n^{\text{mean}}(i) \le F_n^{\text{max}}(i)$. In the case of equality, which only obtains for 'classical' similarity searching where each reference-set molecule is treat independently using a single similarity measure, the test-set molecules lie on the diagonal.

Interpreting multi-fusion similarity maps

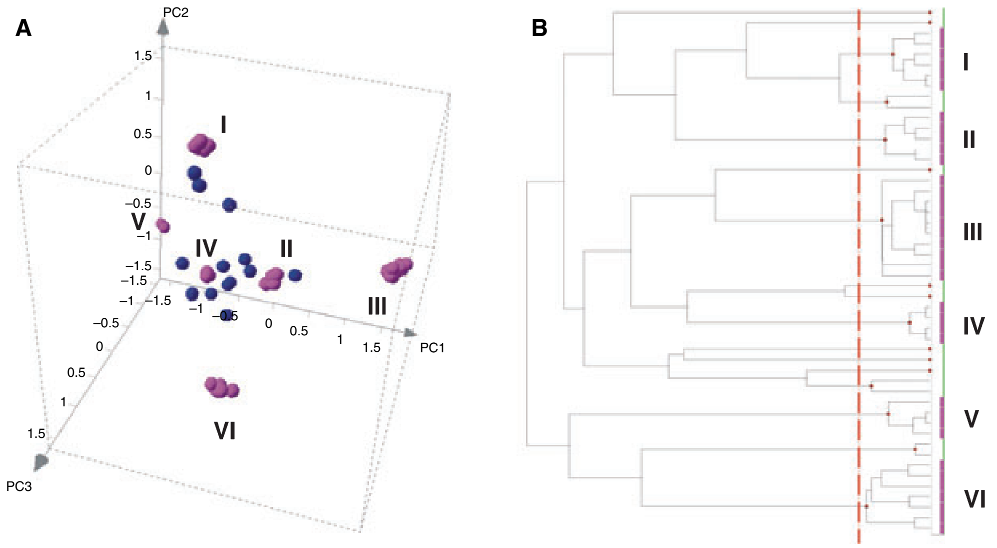
Prototype data set

The fusion-based approach described above is applied to a subset of 50 molecules obtained from the Binding Database (51), a public database that is well suited to the study of structure–activity relationships, to illustrate how MFS maps can be interpreted. The molecules were selected in such a way as to ensure that a variety of molecular scaffolds is chosen from those available in the database. The following protocol is used to generate a chemistry-space. First, the molecular similarities are computed, as described in Section Basis of the multi-fusion approach: generating multi-fusion similarity maps, for the subset of 50 molecules, which has a mean Tanimoto similarity of 0.40 with a standard deviation of 0.20. Secondly, a principal component analysis (9) is carried out taking the similarity matrix as the data matrix (8). Thirdly, as illustrated in Figure 2A, the molecules are displayed in the 3D subspace formed by the first three principal components, which represents about 76% of the variance of the sample. The molecules are also clustered using the complete-linkage hierarchical clustering algorithm implemented in Spotfire[e]; the dendrogram for this clustering is depicted in Figure 2B. The Roman numerals to the right of Figure 2B label six of the clusters in the partition of the data set induced by a molecular similarity value of 0.85, which is represented by the vertical, dashed red line in the figure. Each of the labeled clusters contains at least four molecules and has a minimum intra-cluster similarity of ≥0.87, as shown in the insert table in Figure 2. The six clusters are colored red in the chemical space depicted in Figure 2A.

Interpreting MFS maps with respect to different model reference sets

The 50 molecules of the prototype set are further divided into two subgroups: five molecules are chosen to represent the reference set and the remaining 45 molecules the test set. Each of the molecules in the reference set is also compared with all of the other molecules in the set but not with itself, a procedure called self-referencing that is discussed further in *Case 4*. To illustrate the different features of MFS maps, four choices of the five reference-set molecules are considered. These choices correspond to four cases that clarify a number of the salient features of the proposed approach (vide infra). All of the chemical spaces depicted in Figures 2A, 3A, 4A, 5A and 6A involve the same set of 50 molecules and are, thus, identical, although they may appear somewhat different because of the viewing angles in the figures. For consistency, molecules in the reference set are colored red and those in the test set are colored blue, except for a few that are colored yellow to distinguish them from the bulk of the test set molecules.

Case 1 The reference set of molecules is made up of two molecules from cluster **III**, two from cluster **VI,** and one from cluster **V**. It is clear from Figure 2 that all of the clusters have high intra-cluster similarity and are well separated in chemical space. This is also illustrated in Figure 3A that depicts the chemical space and Figure 3B that depicts the corresponding MFS map. Specific subsets of the test-set molecules are labeled {**a**}, {**b**}, and {**c**} to clearly distinguish them from the remaining test-set molecules in Figure 3. The positions of the test set molecules in Figure 3B depend on



Cluster	No. molecules	Minimum similarity
I	6	0.93
II	5	0.91
III	10	0.90
IV	4	0.96
V	4	0.92
VI	7	0.87

Figure 2: (A) Depiction of the chemical space of the 50 molecules used to describe the prototype MFS maps. (B) Dendrogram depicting the hierarchical clustering of the 50 molecules computed by the complete linkage method[e] using Tanimoto similarity and MACCS key fingerprints (see Section Basis of the multi-fusion approach: generating multi-fusion similarity maps for details). The table at the bottom of the figure summarizes the information on the five major clusters (**I**, **II**, **III**, **IV**, and **V**) that contain four or more molecules with intermolecular similarities of ≥0.9. An additional cluster, **VI**, contains seven molecules with intermolecular similarities of ≥0.87. These clusters are also labeled in the chemical-space plot in (A).

their relationship to the molecules in the reference set, while the position of the reference-set molecules indicates the relationship of the test set molecules to each other (i.e. self-referencing). The tight cluster of reference-set molecules located at a max-fusion value of approximately 0.9 and a mean-fusion value near 0.5 represents the subset of reference molecules that lie near to the clusters of test set molecules denoted by {**b**} in Figure 3A. In contrast, the lone reference-set molecule located at max-fusion and mean-fusion values of approximately 0.2 represents the single reference-set molecule located near {**a**} in Figure 3A.

The subsets in the MFS map distinguished by {**a**}, {**b**}, and {**c**} are associated with the corresponding regions in the chemical space (vide supra) depicted in Figure 3A. These regions are interpreted in the following way: {**a**} corresponds to a pair of molecules that are very similar to the single reference molecule located nearby (high max-fusion value) but far from the remaining four reference compounds (low mean-fusion value); {**b**} corresponds to 13 molecules with max-fusion values of ≥0.90 with respect to the reference molecules that are located in clusters **III** and **VI** of Figure 2A. In addition, as their mean-fusion values are ≥0.54 they are also similar to all of the reference molecules located in these clusters, which is in stark contrast to the molecules in {**a**} that have much lower mean-fusion values. If, for example, the reference set was made up only of active molecules, molecules in the {**b**} test subset would be more likely to be active than those in the {**a**} test subset. This follows since the likelihood that a molecule located within the neighborhood of a set of known actives is also active is greater than the likelihood of a molecule located near to a single, isolated active is active. However, this is not to say that the region of chemical space surrounding the singleton active is not filled with actives, which may very well be the case, but rather than no data exists that supports this contention. Thus, given the data, the active in the latter case could be the result of an assay error or of an improper assignment, whereas in the former case the presence of a number of nearby actives reinforces, but does not prove, the belief that this is, indeed, an 'active region' in chemical space. As in group fusion, MFS maps do not distinguish the intra-cluster similarity of test-set molecules in, for example, {**b**}. As illustrated in this case, the molecules in {**b**} reside in two compact but well-separated clusters, while {**c**} corresponds to a single molecule that is very dissimilar from any of the molecules in the reference set.

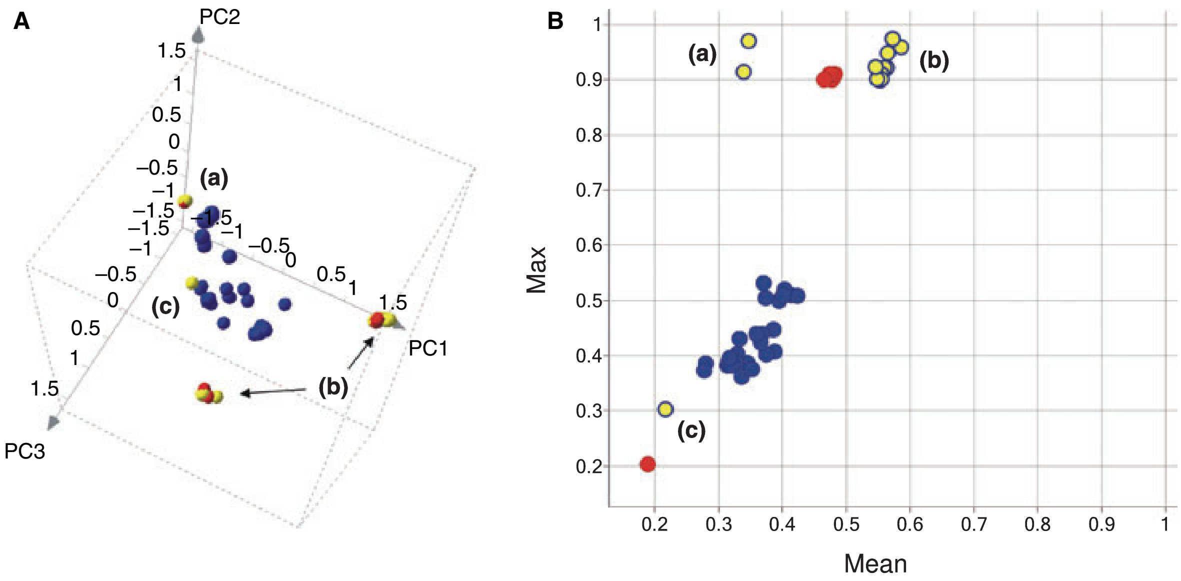


Figure 3: (A) Depiction of the chemical space of the set of 50 molecules illustrated in Figure 2. (B) The corresponding max–mean multi-fusion similarity map. Molecules in the reference set are colored red in both plots. The reference set is made up from clusters well separated in chemical space but with high intra-cluster similarities (see text for details). Test-set molecules are colored blue except for a select few that are colored yellow and comprise groups labeled by {**a**}, {**b**}, and {**c**}.

Case 2 Here, the reference set is composed of a single cluster of five very similar molecules taken from cluster **I** in Figure 2A. The chemical space and corresponding MFS maps are illustrated in Figure 4A,B, respectively. Four of the molecules in the test set are labeled {**a**}, {**b**}, {**c**}, and {**d**} to distinguish them from the remaining 41 molecules in the test set.

As expected, if the group of reference molecules lies in a cluster with high intra-cluster similarity, then the max-fusion similarities are approximately directly proportional to the corresponding mean-fusion values. In the extreme case of a single reference molecule, the max-fusion and mean-fusion values are equal (see Section Basis of the multi-fusion approach: generating multi-fusion similarity

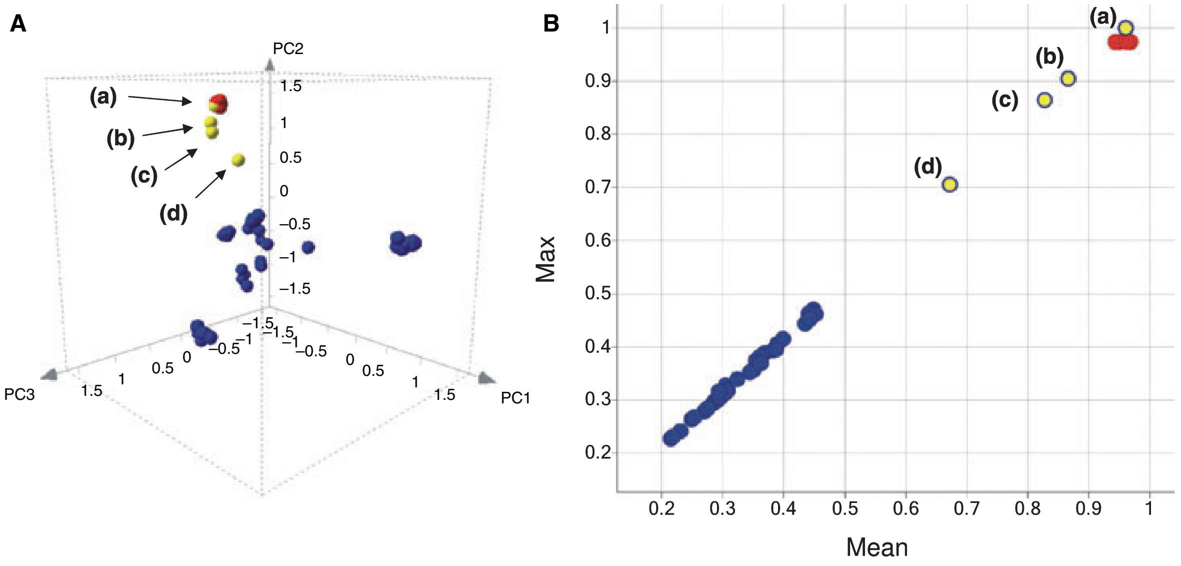


Figure 4: (A) Depiction of the chemical space of the set of 50 molecules illustrated in Figure 2. (B) The corresponding max-mean multi-fusion similarity map. The five reference-set molecules are colored red and form a single, tight cluster in chemical space with a mean similarity of 0.95 and a standard deviation of ±0.02. Test-set molecules are colored blue except for a select few that are colored yellow and labeled by {**a**}, {**b**}, {**c**}, and {**d**}.

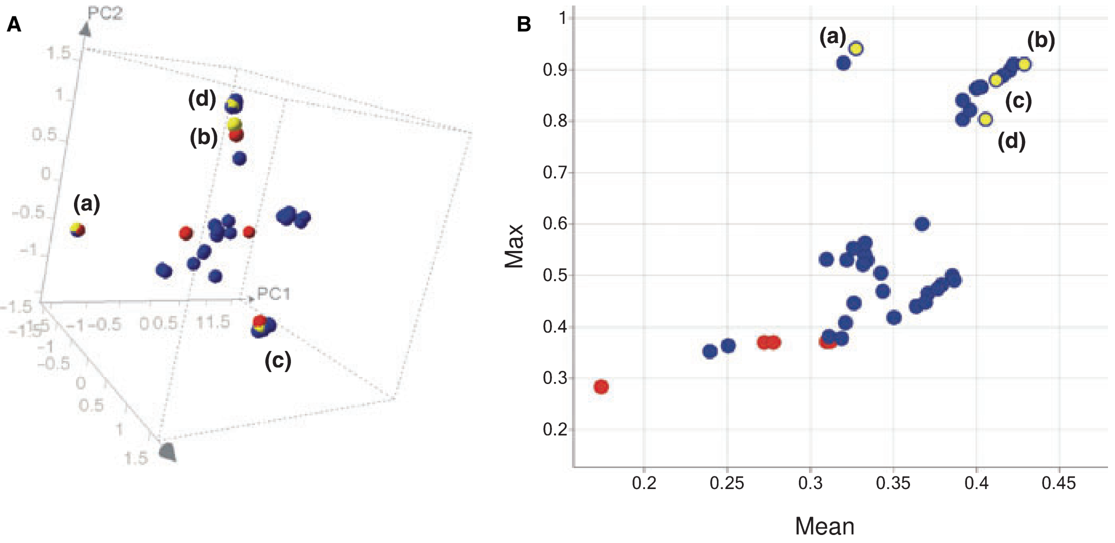


Figure 5: (A) Depiction of the chemical space of the set of 50 molecules illustrated in Figure 2. (B) The corresponding max–mean multi-fusion similarity map. The five reference-set molecules are colored red and represent a diverse group of molecules with a mean similarity of 0.27 and a standard deviation of ±0.1. Test-set molecules are colored blue except for a select few that are colored yellow and labeled by {**a**}, {**b**}, {**c**}, and {**d**}.

maps). Thus, when a single reference molecule is used, as is the case in typical similarity searching methods, nothing is gained using a fusion rule, which is why in similarity fusion studies multiple similarity methods are used (34,52,53).

Case 3 In this case, the reference set is made up of a diverse set of five molecules scattered throughout the chemical space, as depicted in Figure 5A. One molecule was selected from cluster **V** and one from cluster **VI** in Figure 2; the other three molecules were selected from isolated clusters containing one or two molecules. This is in contrast to *Case 1*, where there are two tight, but well-separated, clusters of reference molecules. As before, several of the test-set molecules labeled {**a**}, {**b**}, {**c**}, and {**d**} are distinguished from the remaining test-set molecules. The corresponding MFS map is given in Figure 5B. The self-reference values of the five reference-set molecules are located at the bottom left side of the MFS map. As the reference set is made up of diverse molecules, low max-fusion and mean-fusion values are observed as expected.

The positions of the four test-set molecules corresponding to the singleton subsets {**a**}, {**b**}, {**c**}, and {**d**} in the MFS map in

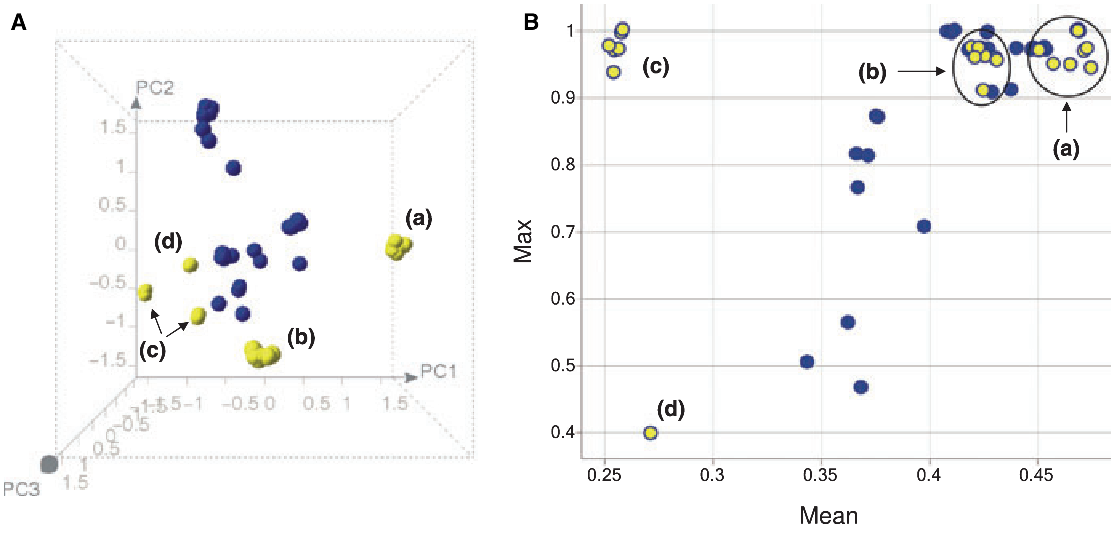


Figure 6: (A) Depiction of the chemical space of the test set of 50 molecules illustrated in Figure 2. (B) The corresponding MFS map for the self-reference case. Selected molecules are colored yellow and labeled by {**a**}, {**b**}, {**c**}, and {**d**}.

Figure 5B, can be interpreted as follows. The molecule in {**a**} is very similar to a single reference molecule ($F^{\max} = 0.94$), but is quite far removed from the remaining reference-set molecules ($F^{\text{mean}} = 0.33$). The three molecules in {**b**}, {**c**}, and {**d**} are progressively less similar to the reference-set molecules as indicated by their steadily decreasing max-fusion values ($F^{\max} = 0.91, 0.88, \text{and}0.80, \text{respectively}$) and their nearly constant mean-fusion values ($F^{\text{mean}} \approx$ 0.42). Interestingly, while the test-set molecules in {**a**}, {**b**}, {**c**}, and {**d**} tend to be clustered in the MFS map, they can be quite separated in chemical space, as illustrated in Figure 5A.

Case 4 This case illustrates self-referencing of the set of 50 molecules, which is accomplished by comparing the set with itself using the same multi-fusion approach described above. The procedure is carried out by comparing each molecule in the set with every other molecule in the set, except itself. The chemical space and corresponding MFS map are illustrated in Figure 6A,B, respectively. To facilitate the discussion, selected groups of molecules are labeled {**a**}, {**b**}, {**c**}, and {**d**} to distinguish them from the bulk of the molecules in the set. Molecules in group {**a**} come from cluster **III** in Figure 2A and have large max- and mean-fusion values ($F^{\max} \geq 0.90$, $F^{\text{mean}} \geq 0.45$). Based on the chemical space of the 50 molecules in the prototype set, it is expected that molecules in group {**b**} in Figure 6A, which come from cluster **VI** in Figure 2A, should also have large max-fusion and mean-fusion values. This is, indeed, the case as illustrated in Figure 6B, although the mean-fusion values are less than those in group {**a**}. On the other hand, molecules in group {**c**}, which are not highly clustered are, nonetheless, grouped together in small groups of approximately two molecules that are relatively well separated from the remainder of the molecules. This should give rise to large max-fusion but much smaller mean-fusion values as can be seen in Figure 6B. The lone molecule in group {**d**} is well separated from the remaining molecules and, thus, should possess both low max- and mean-fusion values, which is the case. As pointed out earlier, clusters of compounds that are separated in the chemical space, such as groups {**a**} and {**b**}, are not necessarily distinguishable in the MFS map.

Although the discussion in this section has focused on the relationship of distributions of molecules in chemical space and their relationship to the corresponding distributions of test-set molecules in MFS maps, the inverse process is also useful. In such cases, it is possible to infer the distribution of molecules in chemical space from the corresponding distribution of test-set molecules in an MFS map, especially when the distribution is simple as is the case when the max-fusion and mean-fusion values tend to be highly correlated.

The examples discussed above are not intended to be definitive but rather as illustrations that should help clarify the way in which chemical spaces can be mapped onto lower-dimension spaces that, nonetheless, are capable of capturing a number of the salient features of the corresponding chemical spaces. However, it must be borne in mind that lower-dimensional representations cannot, in general, capture all of the information of higher-dimensional spaces. Nevertheless, in many instances they can faithfully capture relationships that may be obscured in other, higher-dimensional representations, especially when the higher-dimensional representations are portrayed in two and three dimensions.

Overview of compound databases used in this work

In the present study, four compound collections/libraries (DrugBank, NCIDiv, CNS, and Kappa), with a range of sizes and diversities, are characterized and compared using MFS maps. A summary is presented in Table 2 including the sizes and an assessment of the mean similarity and standard deviation for each data set. Similarity was computed using MACCS key fingerprints and the Tanimoto similarity coefficient, as described in Section Basis of the multi-fusion approach: generating multi-fusion similarity maps. DrugBank, NCIDiv, and CNS data sets are the most diverse with a mean similarity of approximately 0.3. Although all of the compounds in CNS database are CNS active, the compounds are associated with a number of different CNS targets. As expected, the Kappa data set, which is more 'target oriented' has a slightly higher mean similarity of 0.6.

Figures 7 and 8 depict the distribution of molecular weight and logP, respectively, of the compound databases considered in this study. The overall distributions of molecular weight for the DrugBank and NCIDiv databases are similar in shape, as depicted in Figure 7, and possess medians of 325 and 300, respectively. Notably, compounds with kappa activity show a tendency toward increased molecular weight. With respect to logP the DrugBank and

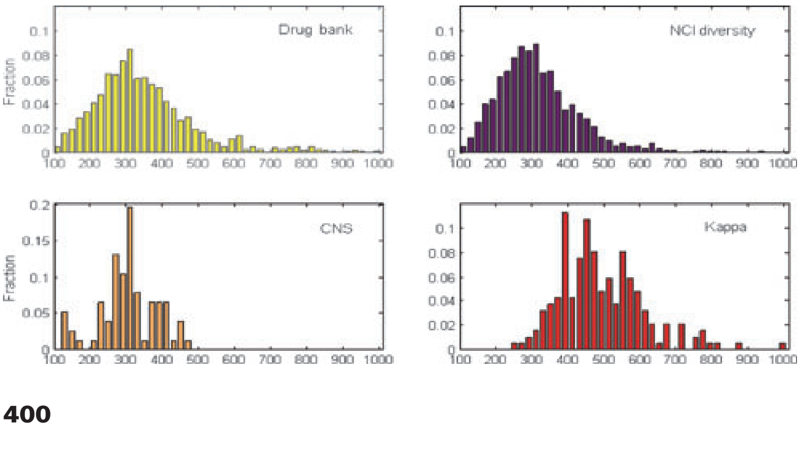


Figure 7: Molecular weight distributions for the four data sets analyzed in this study. The medians of the distributions are, respectively, 325 (DrugBank); 300 (NCIDiv); 307 (CNS); and 471 (Kappa).

NCIDiv data sets have similar distributions with medians of 2.5 and 2.4, respectively, as illustrated in Figure 8. The CNS and Kappa data sets tend to be more hydrophobic, and thus, their logP distributions have medians of 3.4 and 4.3, respectively.

The chemical space for the four compound data sets, depicted in Figure 9, was determined as described for the 50-molecule prototype set in Subsection Prototype data set. The molecules in this and all other chemical-space maps are represented as colored balls of finite radius. Although this improves visualization, it does obscure some of the important spatial relationships among the molecules in chemical space. The most significant effect is that it makes the density of molecules appear much greater than it actually is, and thus, makes it difficult to discern some of the 'natural' clustering present in almost all chemical spaces. Figure 10 presents the same view of the chemical space illustrated in Figure 9, the only difference being that the molecules are represented as colored points rather than balls. Although the latter figure provides a more accurate depiction of the sparse and clustered distribution of molecules typically seen in chemical spaces, it is generally not used here because it can be difficult too 'see' some of the molecules. Thus, even though the colored ball representation is used throughout this paper, it is important to realize that chemical space is typically quite sparse and tends to be populated with many small clusters of molecules.

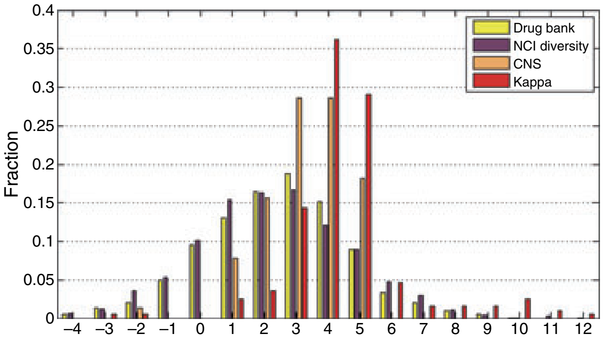


Figure 8: LogP distributions for the four data sets analyzed in this study. The medians of the distributions are, respectively, 2.5 (DrugBank); 2.4 (NCIDiv); 3.4 (CNS); and 4.3 (Kappa).

Results and Discussion

General overview

The following subsections will cover comparison of compound collections, compound selection and acquisition, ligand-based virtual screening, and computationally based applications of MFS. As in the current work emphasis is placed on the visual characterization and comparison of actual compound libraries and collections, the dynamic visualization capabilities of the Spotfire™ program[e] are

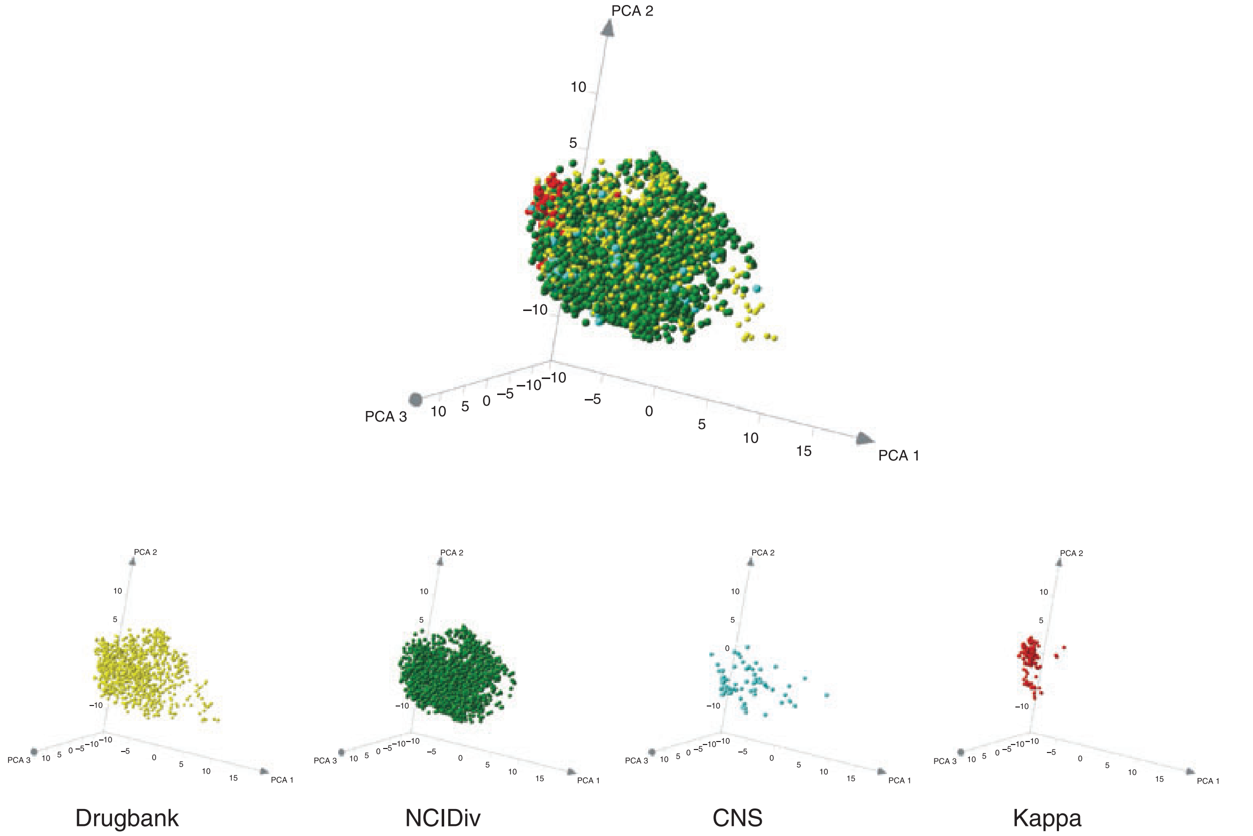


Figure 9: Depiction of the chemical spaces of the databases considered in this study: DrugBank, NCIDiv, CNS, and Kappa. The four insets shown below depict the chemical spaces covered by each of the individual databases.

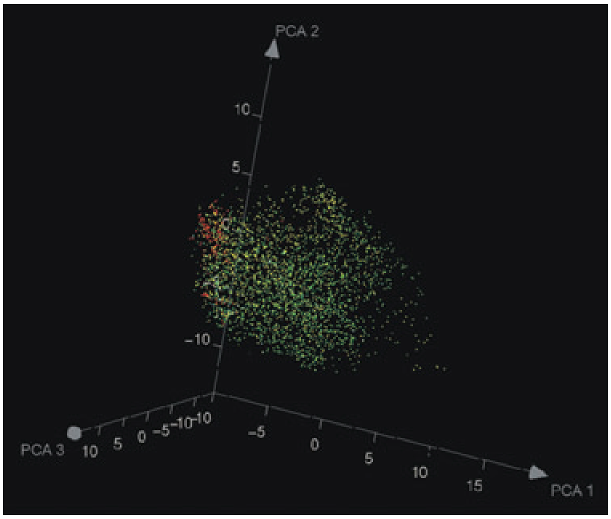


Figure 10: A more realistic depiction of the density of the chemical space of the four data sets considered in this work: DrugBank (yellow); NCIDiv (green); CNS (cyan); and Kappa (red).

quite useful in the analysis. This is not to say that the analysis cannot be carried out effectively without the use of Spotfire™, only that it is facilitated by the use of such dynamic display capabilities. Any program with similar capabilities will be quite suitable for carrying out the analysis described here. Moreover, programs without such dynamic display capabilities can still be used, but the analysis is considerably more cumbersome.

Three-dimensional representations of chemical space are ubiquitous in cheminformatics and they can provide a considerable amount of useful qualitative information regarding the nature of large or small sets of compounds, which includes NN relationships, compound clustering, and molecular diversity. However, the magnitude of units of measure associated with each of the co-ordinate axes are difficult to relate to notions of molecular similarity, and thus, the relationship between the apparent proximity of compounds in a chemical space and their molecular similarity can be difficult to ascertain visually. In contrast, the max-fusion and mean-fusion similarity values are given explicitly on the MFS maps (vide infra). As will be seen in the sequel, MFS maps provide another window into chemical space that gives rise to powerful synergies when used in conjunction with traditional 3D maps of chemical space. This is illustrated in Figure 11. Figure 11A shows the chemical space of two test sets of compounds derived from combinatorial libraries, {**a**} and {**b**}, colored blue and yellow, respectively, each with 100 molecules. The reference set, colored red, is the set of CNS-active

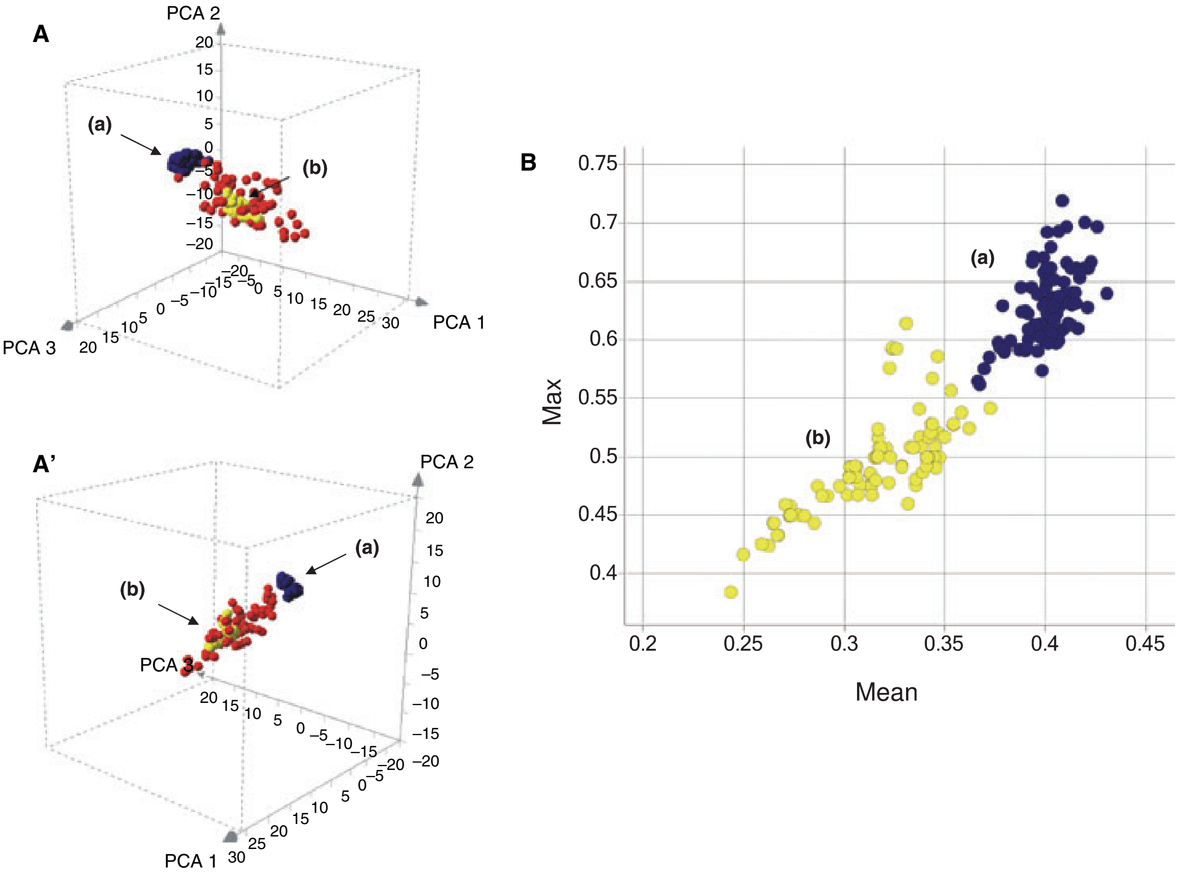


Figure 11: (A and A′) Orthogonal views of the chemical space of two sets of 100 compounds, {**a**}, and {**b**}, colored blue and yellow, respectively, and the CNS-active compounds, colored red. (B) The corresponding max–mean MFS map with the CNS-active compounds as the reference set.

compounds. From Figure 11A, it appears that the molecules in {**a**} generally lie further from those of the reference set than do the molecules of {**b**}. This relationship holds true even in the orthogonal view as illustrated in Figure 11A′. Figure 11B, however, tells a different story. From the discussions in Sections Basis of the multi-fusion approach: generating multi-fusion similarity maps and 2.2, it follows that a significant majority of the molecules in {**a**} lie closer to the molecules in the CNS reference set than do those of {**b**}, as they possess both larger max-fusion and mean-fusion similarities. Thus, it is clear from this example that visual analysis of chemical spaces alone can lead to incorrect inferences regarding the nature of the chemical space under consideration. Importantly, as will be seen in the following examples, it is the synergies produced by combining the information in the chemical-space and MFS maps that yields the most significant insights.

Visualization methods obviously have limitations, especially when dealing with large collections of compounds. As discussed in Section Computationally based applications of the multi-fusion similarity method, the present approach can also be applied in a purely computational manner, using well-developed multi-criterion decision-making (MCDM) methods based on multi-objective optimization techniques. These methods have been applied in a wide variety of business applications (54) as well as in a number of combinatorial library design applications (55–58).

Comparing compound collections

Two examples are presented in this section that illustrate how the MFS approach can be applied in comparing compound collections. The examples are based on the four compound collections/libraries described in Section Overview of compound databases used in this work – DrugBank, NCIDiv, CNS, and Kappa. The first example involves a comparative study of DrugBank and NCIDiv. Figure 12 depicts four MFS maps: the reference sets are designated along the top and the test sets along the left-hand side of the figure; DrugBank molecules are colored yellow and NCIDiv molecules are colored red; the two maps along the main diagonal are self-referential. Identical molecules have been removed in self-reference maps so that the only molecules with max-fusion similarity values of unity are either stereoisomers or non-identical molecules that are not resolved by the MACCS key fingerprints. In the MFS map in the upper right corner, NCIDiv is the reference set and DrugBank the test set, while the opposite is true for the MFS map in the lower left corner of the figure. In contrast to the self-reference case, molecules with max-fusion similarity values of unity located along the

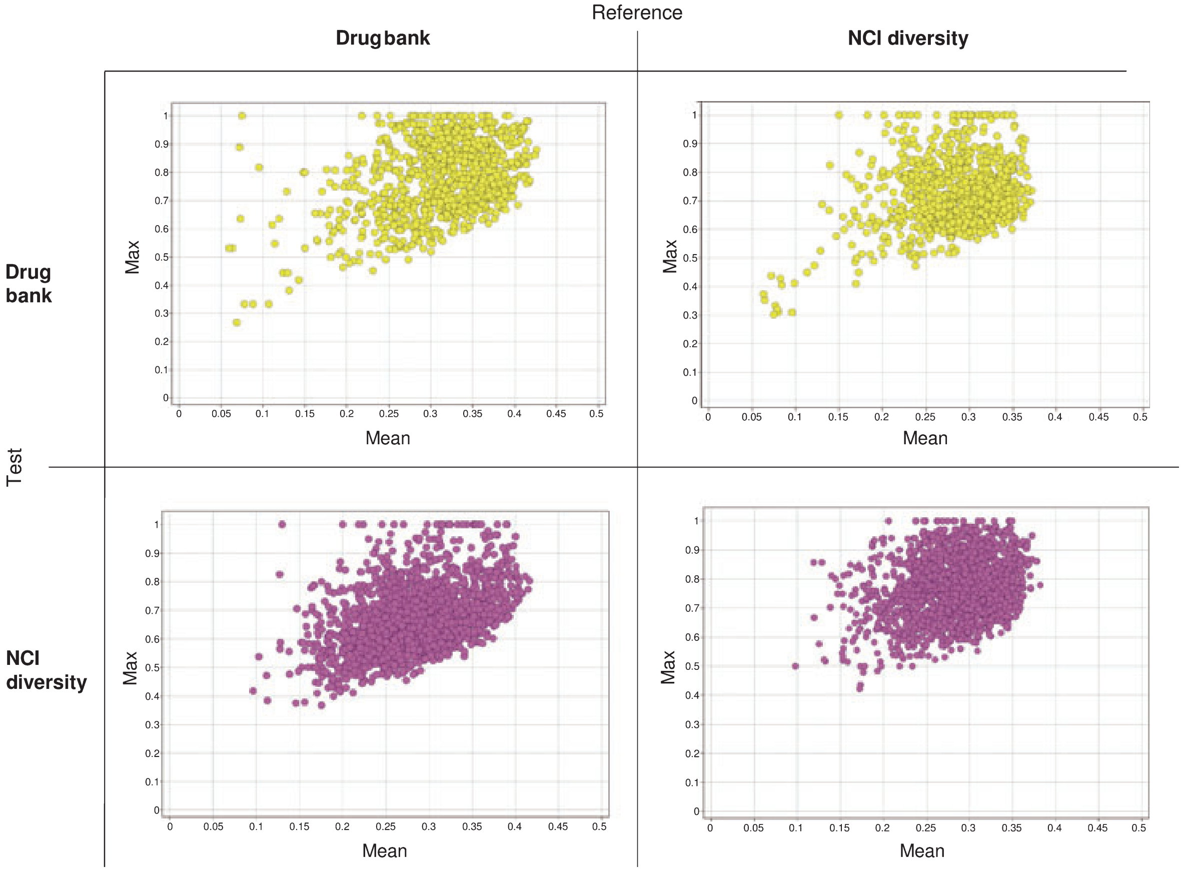


Figure 12: Multi-fusion similarity maps comparing the DrugBank and NCIDiv compound collections, colored yellow and red, respectively. The two plots along the principal diagonal (upper left to lower right in the figure) correspond to self-referencing MFS maps of the two compound collections.

top of each of the off-diagonal maps may be identical, indicating that the same molecule is present in each of the data sets. All of the molecules in the maps shown in the figure lie well above the line that represents equality between max- and mean-fusion similarity values (see Figure 1). As discussed in Section Interpreting multi-fusion similarity maps, such cases indicate the presence of tight clustering within the reference set, a situation that is not expected in DrugBank and NCIDiv compound collections, especially in the latter case. The top two maps in Figure 12 indicate that some of the molecules in DrugBank lie somewhat removed from the bulk of DrugBank molecules regardless of whether DrugBank is referenced to itself or to the NCIDiv collection. That this is, indeed, the case can be seen by considering the chemical space depicted in Figure 9. By comparison, the NCIDiv data set appears to fill chemical space in a more uniform manner as evidenced by the more compact distributions shown in the bottom maps of Figure 12. Nevertheless, it does appear that DrugBank occupies some regions of chemical space not occupied by NCIDiv. Thus, as will be shown in Section Compound selection and acquisition, molecules from DrugBank could be used to augment the NCIDiv data set in a way that increases its coverage of chemical space and, hence, its diversity.

The second example considers the relationship of a given test set – DrugBank – to two small, reasonably focused libraries – Kappa and CNS – taken as reference sets. Figure 13 depicts the chemical space and MFS maps for the Kappa/DrugBank case. Figure 13A,B shows two, orthogonal chemical-space views, where molecules in the Kappa reference set are colored red and those in the DrugBank test set are colored yellow. The compounds colored blue and labeled {**a**}, {**b**}, and {**c**} represent specific subsets of test molecules. It is clear from the chemical-space maps that each of the subsets lies in different regions of chemical space with respect to the Kappa reference set. The corresponding MFS plot is given in Figure 13C. The subsets of test set molecules colored blue are evident in this figure. The molecules in {**a**} possess the largest max- and mean-fusion similarity values indicating that they lie close to a large cluster of molecules in the Kappa reference set. In contrast, the molecules in {**b**} possess comparable max-fusion similarity values to those in {**a**} but possess smaller mean-fusion values indicating, as is clear from Figure 13A,B, that they lie within a chemical-space region of the Kappa reference set that is less clustered than the molecules in {**a**}. The molecules in {**c**} have the lowest max- and mean-fusion similarity values of any of the molecules in the test set indicating that they are far removed from any of the reference set molecules. The remainder of the molecules in the test set shows an approximately linear relationship indicating that the Kappa reference set molecules mostly lay in a single, large cluster in chemical space, as observed in the figures.

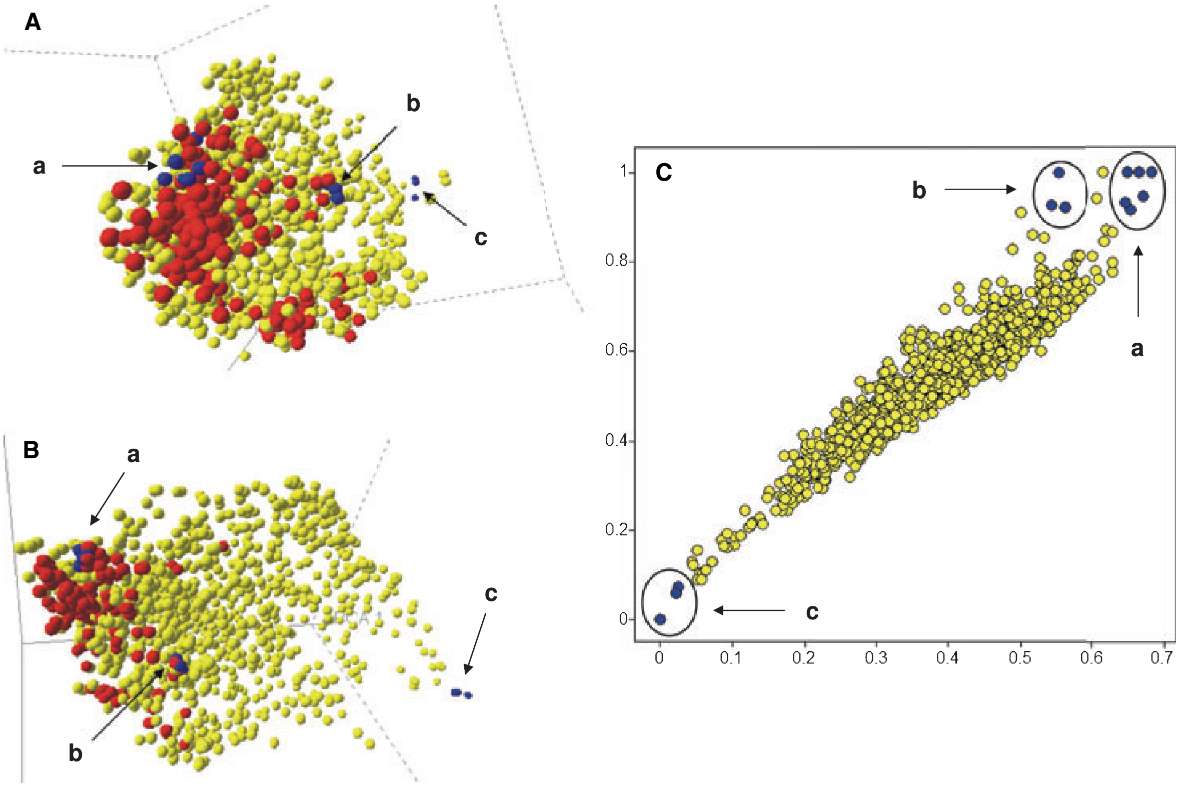
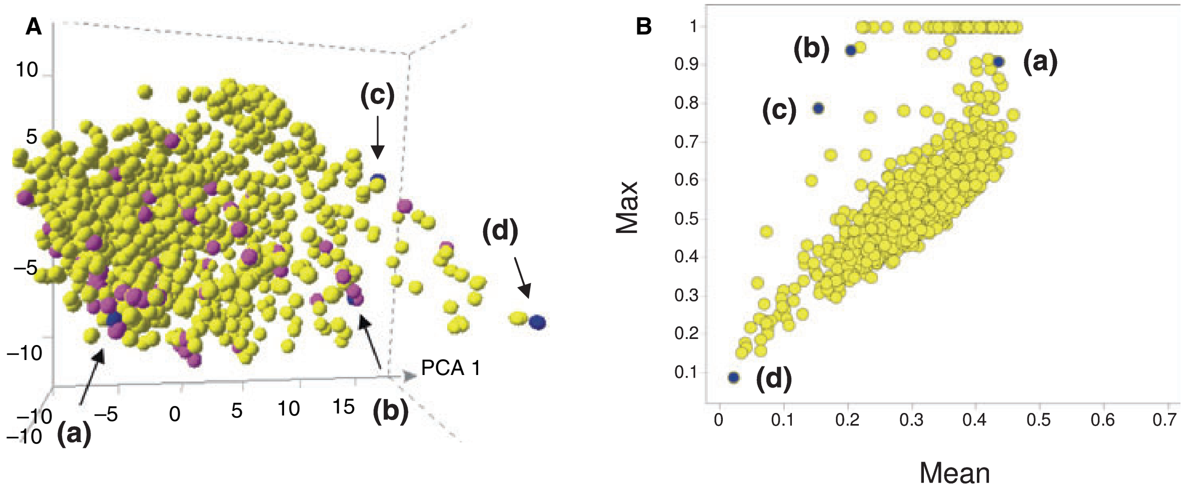


Figure 13: Comparison of 1055 test-set molecules from DrugBank, colored yellow, to the 196 reference-set molecules from Kappa, colored red. (A and B) Correspond to two orthogonal views of the chemical space. (C) Depicts the corresponding MFS map, where reference set molecules have been omitted. As can be seen in (A, B, and C) DrugBank test-set molecules in group {**a**} lie very close to several Kappa reference-set molecules, test-set compounds in group {**b**} lie close to fewer reference-set compounds, and test-set molecules in {**c**} are removed from essentially all of the molecules in the reference set.

In static figures such as the maps depicted in Figure 13, it is difficult to gain a sense of the power of Spotfire™ to dramatically assist the process of visual exploratory data analysis (EDA). For example, by selecting the molecules colored blue lying within the three ellipses in Figure 13C, their positions are immediately identified in the chemical-space maps given in Figure 13A,B. This process is further facilitated by the fact all three maps appear on the same screen simultaneously. The process can, of course, also be reversed: molecules selected in the chemical-space representation are immediately identified in the MFS plot. The interplay between the chemical-space and MFS representations is facilitated by the dynamic display capabilities of Spotfire™ that significantly enhance visual EDA. In addition, once molecules have been identified in any plot, it is easy to retrieve associated information including, for example, their structures and physico-chemical properties. The input data may contain SMILES strings, CAS numbers, names, ID numbers, etc. The automatic display of names using Spotfire is illustrated by the table in Figure 14.

The third example in this section deals with the CNS and DrugBank compound collections. The chemical space and MFS maps are given in Figure 14A,B, respectively. Molecules in the CNS reference set are colored red in Figure 14A; DrugBank test-set molecules are colored yellow in Figure 14A,B. The blue test-set molecules, labeled {**a**}, {**b**}, {**c**}, and {**d**}, correspond to ethylthioperazine, tranylcypromine, valproic acid, and calcium chloride, respectively, as shown in the table. The four molecules were selected from Figure 14B based on their widely different positions in the MFS plot. While molecules {**a**} and {**b**} both have high max-fusion similarities, indicating that they are both close to at least one member of the CNS reference set, {**a**} has a significantly greater mean-fusion similarity, which indicates that it is closer to a cluster of reference molecules than {**b**}. Molecules {**c**} and {**d**} are even more isolated, with {**c**} having slightly more proximity to the CNS reference set. The group of yellow test-set molecules lying across the top of Figure 14B at a max-fusion similarity value of 1.00 is made up either of molecules structurally identical to molecules in the CNS reference set or, due to the limitations of the MACCS key fingerprint representation, appear to be made up of molecules identical to CNS reference-set molecules. There are 59 DrugBank molecules with max-fusion similarities of unity, but just one is a stereoisomer of a CNS drug (zopiclone – eszopiclone); 58 molecules are identical to CNS-active molecules.

From the above discussion and based on the way they were generated, it is not unexpected that the Kappa and CNS reference sets



Selected data points:

Library	Substance ID	Name	
Drugbank	9342	Ethylthioperazine	**(a)**
Drugbank	9364	Tranylcypromine	**(b)**
Drugbank	9394	Valproic acid	**(c)**
Drugbank	10330	Calcium chloride	**(d)**

Figure 14: Comparison of 1055 test-set molecules from DrugBank, colored yellow, to the 77 reference-set of CNS-active molecules in the CNS database, colored red. (A) Depicts the chemical space and (B) the corresponding MFS map. The reference set has been omitted from the MFS map. Selected DrugBank test-set molecules denoted by {**a**}, {**b**}, {**c**}, and {**d**}, colored blue correspond to ethylthioperazine, tranylcypromine, valproic acid, and calcium chloride, respectively.

should possess somewhat different chemical-space characteristics: the Kappa reference set contains molecules that are active against a single target, namely the κ opioid receptor. The CNS reference set, on the other hand, contains molecules that are active against one or more of a variety of CNS targets (see Section Overview of compound databases used in this work for additional discussion). Thus, the CNS reference set should be and is more diverse than the Kappa reference set. This is also clear from the 3D chemical-space maps in Figures 13A and 14A, where the Kappa and CNS-active molecules are colored red.

Figure 15 shows a cross-comparison of the DrugBank and NCIDiv test sets with respect to the same Kappa and CNS reference sets. The data in Figure 15A,B was presented earlier in Figures 13C and 14B, respectively, and is included here to facilitate comparison with the related data on the NCIDiv test set. Figure 15A,C provides a comparison of the DrugBank and NCIDiv test sets with respect to the Kappa reference set; Figure 15B,D provides the corresponding test set comparison with respect to the CNS reference set. In the former, it is clear that most of the molecules in both test sets bear a similar relationship to the molecules of the Kappa reference set. However, the upper right-hand corner of the DrugBank test set is dramatically different from that of NCIDiv test set (compare Figure 15A,C). This indicates that DrugBank contains a number of molecules that are quite similar to those in the Kappa reference set, while the NCIDiv test set does not. This result is not surprising given the way in which the different libraries were constructed (vide supra).

In the latter, as illustrated in Figure 15B,D, the distribution of points is more spread out in both cases. This is entirely understandable since, as noted earlier, the CNS reference set is much more diverse than the Kappa reference set. In addition, there are a significant number of identical or very similar molecules in DrugBank compared to the NCIDiv set with respect to the CNS reference set. Again, this is not entirely surprising as DrugBank contains some drug molecules with CNS activity that would likely be identical or highly similar to the set of CNS active molecules in the reference set.

Compound selection and acquisition

High-throughput screening (HTS) methods play an important role in modern drug-discovery research. To be effective, these methods require access to large, diverse compound collections. This raises two important issues. The first is related to the optimal selection of diverse subsets for screening (59) from a large compound collection ('dissimilarity sets'), and the second is related to expanding the overall size and diversity of an existing compound collection (19). In a general sense, both problems are very similar. In compound

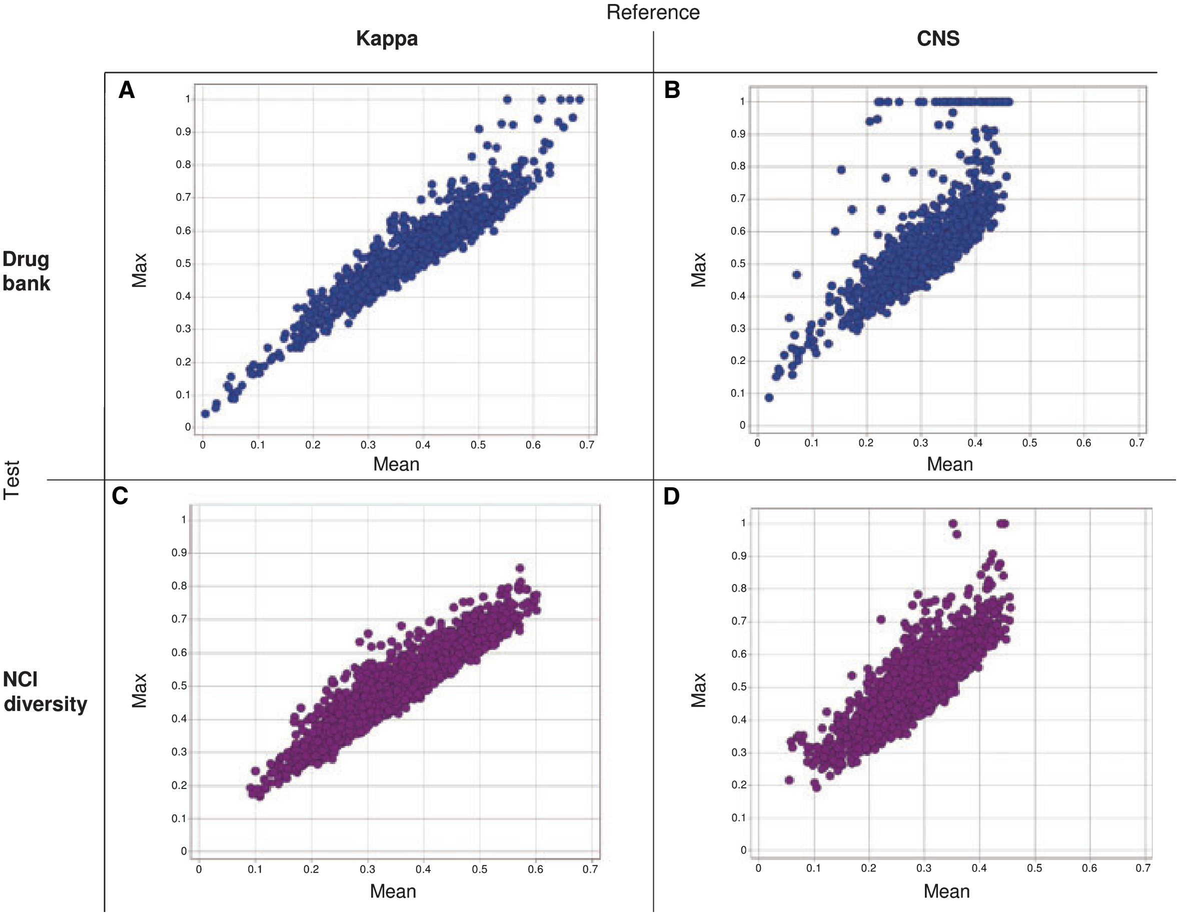


Figure 15: Max–mean MFS maps comparing the DrugBank and NCIDiv test sets to Kappa and CNS reference sets.

selection, a diverse subset is obtained by selecting compounds from a given compound collection. During the process, compounds are added to the growing dissimilarity set, which is initially a null set, in a manner that maintains, in some optimal way, the overall diversity of the growing set. In compound purchasing, the process is basically the same except that compounds are selected from a vendor collection and are added to an existing compound collection, in a manner that maintains, in some optimal way, overall diversity.

Most of the procedures that are applied in either case are designed to select compounds that increase the overall diversity of the nascent set. For a variety of reasons, selecting compounds that are largely isolated from the other compounds within a set (i.e. 'singletons') can have potentially undesirable side-effects. Undoubtedly, there are many ways to address this problem. As seen in the following example based on compound acquisition, the MFS approach may provide an additional means for dealing with this important issue. In this example, NCIDiv is taken as the reference set and DrugBank the test set, but the reverse could also have been done. Figure 16A,B shows two orthogonal views of the underlying chemical space. Molecules in the NCIDiv reference set are colored red and those in the DrugBank test set are colored yellow. The molecules colored blue are selected from DrugBank and occupy four different subsets labeled {**a**}, {**b**}, {**c**}, and {**d**} that cover different regions of chemical space. As seen from the chemical-space representations in Figure 16A,B, the molecules in {**a**} are clearly dissimilar from the remainder of the molecules in either DrugBank or NCIDiv. Not surprisingly they are also found in the lower-left corner of Figure 16C, isolated from the remainder of the DrugBank test-set molecules. Adding molecules from this set to the NCIDiv reference set will certainly increase its diversity. In contrast, the DrugBank molecules in {**c**} are minimally dissimilar from molecules in the NCIDiv reference set and, as such, they are located in the upper right-hand corner of the MFS plot in Figure 16C. In this case, adding these molecules to the NCIDiv reference set will not increase the diversity of the reference set. Analysis of subsets {**b**} and {**d**} is not as straightforward. It is not clear from the two chemical-space representations depicted in Figure 16A,B which

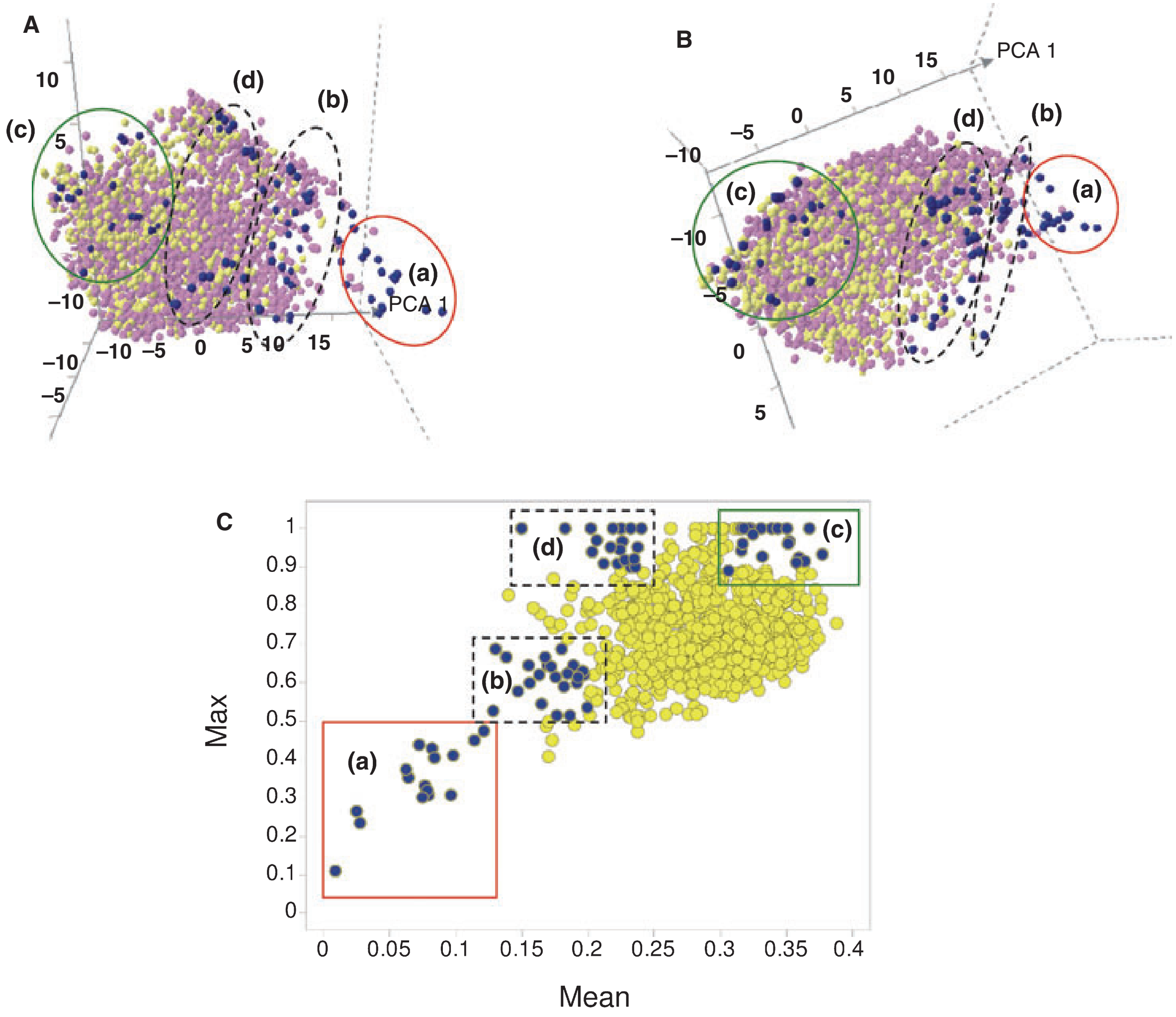


Figure 16: Comparison of the 1055 test-set molecules from DrugBank, colored yellow, to the 1990 test-set molecules from NCIDiv, colored red. (A and B) Depict two orthogonal views of the chemical space, and (C) depicts the corresponding MFS map. Four subsets of DrugBank test-set molecules, colored blue, are indicated by {**a**}, {**b**}, {**c**}, and {**d**}.

molecules should be selected next to increase the diversity of the reference set. It becomes much clearer if one considers Figure 16C. In this case, subset {**b**} would be the best choice. The location of these molecules with respect to chemical space, however, is not straightforward. The last subset, {**d**}, represents the most subtle case. Molecules in this subset cannot be spotted easily, if at all, using chemical-space representations alone. From Figure 16C it is seen that the molecules in {**d**} have high max-fusion similarities with respect to the NCIDiv reference set. However, they also have relatively small mean-fusion similarities. This indicates that the molecules in {**d**} lie in a relatively sparse region of the reference space. Thus, it may be desirable to choose molecules from this set even though they have high max-fusion similarities because their selection will increase the population in a sparse region of chemical space. As noted earlier, it is desirable that the chemical space of a given compound collection not possess too many sparse regions populated with singletons or doubletons.

Admittedly, the graphical approach has limitations when the compound collections become large, a situation that is likely to occur in many cases of compound selection and acquisition. However, as is discussed in Section Computationally based applications of the multi-fusion similarity method, computational implementation of the strategies described here is possible and is the subject of on-going research in our group.

Ligand-based virtual screening

Ligand-based virtual screening has become an important part of the drug-discovery process. As discussed in the Introduction, numerous methods designed to improve the virtual screening process have been developed. The Willett laboratory at the University of Sheffield, in particular, has developed a number of improved virtual screening procedures based on data-fusion methods originally developed for a variety of engineering and data retrieval applications (38,39). In the area of compound retrieval, one of the most successful has been the group fusion method, where it has been shown that max-fusion similarity scores are quite effective in identifying known active compounds (35,37). In contrast, the multi-fusion approach proposed here uses both max-fusion and mean-fusion similarity values to identify potentially active of compounds. This raises the question as to whether the information gained by including two sets of fused similarity values provides any additional benefit. That this may, in fact, be the case follows from the supposition, discussed in Section Interpreting multi-fusion similarity maps, that molecules in close proximity to groups of actives have a greater likelihood of also being active than molecules in close proximity only to singleton actives. As noted earlier, the region surrounding a singleton active may be rich with actives, but whether this is true cannot be ascertained from the existing data alone and therefore cannot be used as a basis for inference.

Figure 17 provides an illustrative example, where the Kappa actives are taken as the reference set and DrugBank is taken as the test set. Figure 17A depicts the chemical space – Kappa reference set molecules are colored red and DrugBank test set molecules are colored yellow. As is clear from the figure, the Kappa reference set occupies a reasonably localized region of chemical space. Molecules colored blue are test set molecules with high max-fusion and high mean-fusion similarity values; they also lie within the box at the upper right-hand corner of the MFS plot in Figure 17B. As indicated in the figure, molecules within the box are all known opioid compounds.

If only max-fusion similarity is used, as in the group fusion approach, all of the molecules in the DrugBank test set will be projected onto the Kappa (Max) axis. This ensures that all of the molecules with high max-fusion similarities will lie at the top of a list ordered by max-fusion similarity values. However, if mean-fusion similarity is used, molecules in the test set will be projected onto the abscissa. As molecules in the test set located near actives on the fringe of the Kappa reference set possess mean-fusion similarity values that tend to be less than those test-set molecules located in the center of the Kappa-active region, their positions in Figure 17B will tend to move toward the left of the figure. When the 'molecular points' are projected onto the mean-fusion similarity axis the ordering of the test set molecules will differ from that obtained using max-fusion similarity. Thus, molecules located at the top of the 'max-fusion' list will tend to move downwards in the 'mean-fusion' list, leading to a decrease in rank correlation. When such a case occurs, more molecules will have to be sampled to ensure that all of the molecules with high max-fusion similarity will also be included in the mean-fusion similarity list. This provides an explanation for the observation that max-fusion similarity generally outperforms mean-fusion similarity as a means for effectively identifying 'unknown' active compounds in ligand-based virtual screening applications (35,37).

As discussed in Section Interpreting multi-fusion similarity maps for *Case 2*, when the reference set of active molecules is grouped into a single, reasonably tight cluster, max-fusion and mean-fusion similarity values are approximately related by a constant. In such cases, either fusion measure would yield the same results. However, as the diversity of the active (reference) set increases the data points will tend to spread out along the mean-fusion axis, even when there is a degree of clustering in the active set. In this case, the max-fusion approach would be expected to yield better results than mean-fusion (vide supra). Although this is not proof, it does provide a plausible 'mechanistic' explanation for the observations made by the Willett group, which are based on comprehensive studies of ligand-based virtual screening of the group fusion procedure (35,37).

An additional aspect of the multi-fusion approach that is not exploited in this section is the ability of the mean-fusion similarity values to further 'spread out' the subset of molecules with high max-fusion similarity values. This provides an additional criterion for identifying potential active molecules since, as discussed in Section Basis of the multi-fusion approach: generating multi-fusion similarity maps and illustrated in Figure 3, molecules with high max-fusion and high mean-fusion similarity values tend to be located in regions containing multiple actives, and thus, are more likely to also be active than molecules located in regions of high max-fusion but low mean-fusion similarity (vide supra). This additional degree of resolution can be of use in cases where there are restrictions, for whatever reason, on the

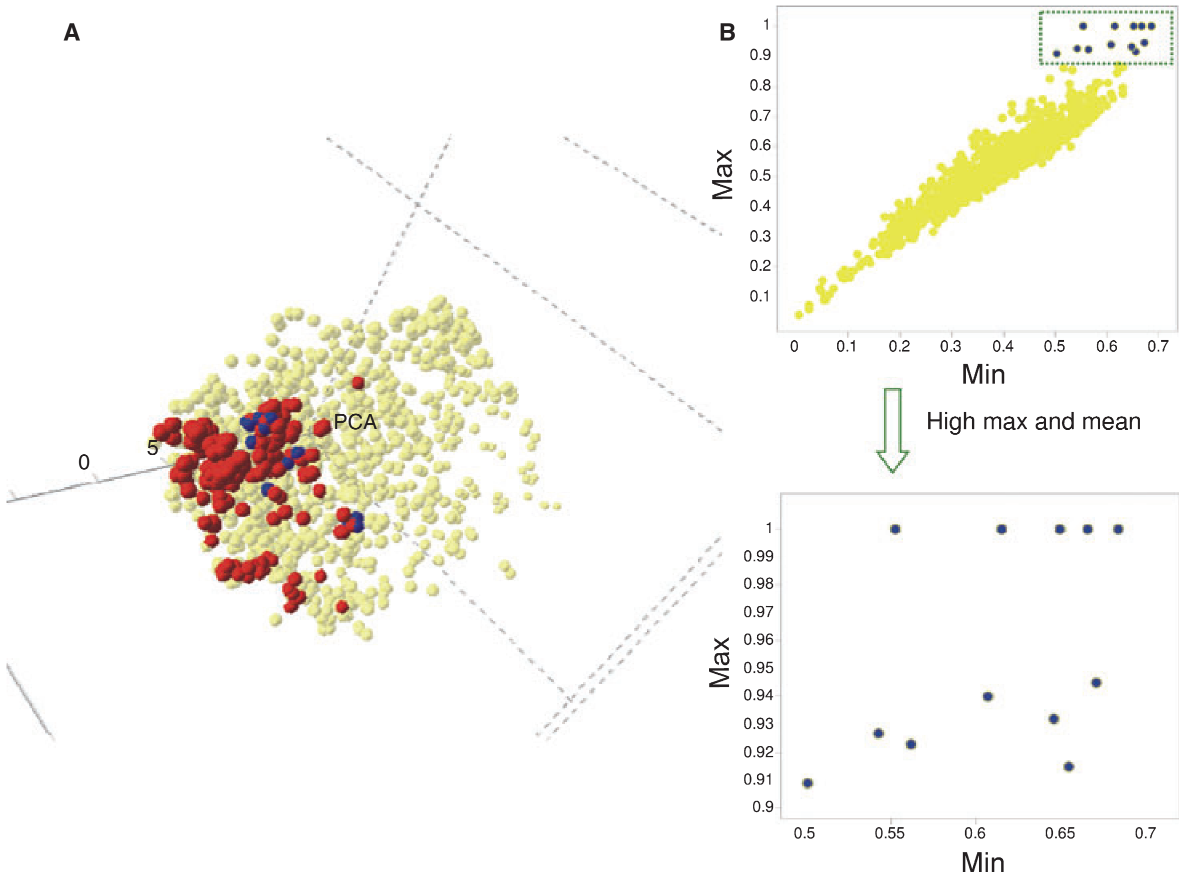


Figure 17: (A) Depicts the chemical space and (B) the corresponding MFS map for the 1055 test-set molecules from DrugBank, colored yellow, and the 77 reference-set molecules from Kappa. Selected test-set molecules, colored blue, lie within the box in the upper right-hand corner of the MFS map in (B). The inset located below it provides a more detailed designation for the 12 blue, test-set compounds.

number of compounds that can be screened in subsequent iterations of a screening campaign. This is not the case in Figure 17B, which shows a relatively modest spread (approximately 0.2 'similarity units', see inset in Figure 17) along the mean-fusion similarity axis. In fact, all of the opioid compounds shown in the figure could have been identified without the need to consider the mean-fusion similarity values in any way.

Computationally based applications of the multi-fusion similarity method

Because the current approach emphasizes data visualization it is not entirely appropriate for drawing detailed conclusions from very large compound collections (1,60) as may be encountered, for example, in typical compound selection and ligand-based applications. Fortunately, some of the computational methods developed for MCDM are also applicable to this problem (54). In particular, a number of methods based on evolutionary algorithms have been developed for tackling the multi-objective optimization problem (61). Applications of multi-objective optimization in combinatorial chemistry for the design of diverse compound libraries with optimum physico-chemical and biopharmaceutical properties have also been reported (55–58).

In the current approach, Pareto optimal solutions for two objective functions based on max-fusion and mean-fusion similarity are determined. This is accomplished by sequentially determining the Pareto optimal subsets of non-dominated solutions. While the non-dominated subsets are ordered with respect to each other, the molecules *within* each non-dominated subset are not. This results in a coarse-grained ordering from the best subset of non-dominated solutions to worst subset of non-dominated solutions. A general description of this approach is clearly presented in the excellent book by Deb (61), which should be consulted for details. The graphical representations described here can then be used to view the results, which will provide an intuitive picture of the solutions obtained from the computations. Work on the computational aspects of the multi-fusion-based similarity is on-going in our laboratories, and the results will be presented in future publications.

Conclusions

The current work describes a novel method for graphically depicting information related to the chemical space properties of compound collections and libraries. The method is based on the use of two-dimensional MFS maps generated from fusion-based molecular similarities. Each of the points in a map corresponds to a test-set molecule whose position is determined by the value of its mean-fusion similarity, which lies along the abscissa, and its max-fusion similarity, which lies along the ordinate of the map. Max-fusion represents the largest computed similarity value of a test-set molecule with respect to any of the molecules within the reference set, while mean-fusion is the average of the computed similarity values of the test-set molecule with respect to all of the reference-set molecules. Because, as seen in eqn 5, max-fusion similarity values are always greater than or equal to their corresponding mean-fusion similarities all of the points in an MFS map lie in the upper triangular region illustrated in Figure 1. Reference set molecules need not correspond only to active molecules; rather, they can represent any set of molecules that is compared to a given test set of molecules. For example, reference sets can be made up from molecules of the test-set itself (the self-referencing case), from molecules of a small library or large compound collection, or from molecules that are active in a given assay or group of assays. Examples, discussing each of these possibilities were presented in Section Results and Discussion. An important feature of MFS maps is that they provide information on the chemical space of each test-set molecule induced by the set of reference molecules, but information on the reference-set molecules themselves is not expressed explicitly in an MFS map.

The emphasis in this work is on the use of multiple fusion-based similarity measures as a basis for representing high-dimensional chemical-space information on compound libraries and collections in graphical form using MFS maps. Several examples were presented illustrating how this can be accomplished and how the graphical representations obtained can be interpreted. It is also shown that in a number of cases there is a synergistic relationship between the two fusion-based similarities so that use of both together provides more information than use of either does separately.

While useful, purely graphical analysis of multiple fusion-based data encounters significant difficulty as the size of the compound libraries and collections being analyzed become very large. However, as discussed in Section Computationally based applications of the multi-fusion similarity method, computational methods based on multi-objective optimization methods employing evolutionary algorithms are available for dealing with this issue, and this is an area where we are directing our future research efforts. Although the methodology described in this work is focused on applications to small molecules, it can be applied to any sets of objects (e.g. proteins) for which a similarity measure can be determined, computationally or otherwise, considerably extending the range of possible applications that can be treated using this methodology.

Acknowledgments

JLM-F and GMM would like to thank the BIO5 Institute at the University of Arizona, Tucson, Arizona and the Torrey Pines Institute for Molecular Studies, La Jolla, California for partial support of this work. RAH would like to thank the National Institute on Drug Abuse (1 R21 DA019620-01A2) and the Arthritis and Chromic Pain Institute, San Diego, California for partial support of this work. The authors are grateful to Dr Tudor Oprea for providing the WOMBAT data.

Reference

1. Scior T., Bernard P., Medina-Franco J.L., Maggiora G.M. (2007) Large compound databases for structure-activity relationships in drug discovery. Mini Rev Med Chem;7:851–860.
2. Leach A.R., Gillet V.J. (2007) An Introduction to Chemoinformatics (revised edition). New York: Springer.
3. Baldi P. (2005) Chemoinformatics, drug design, and systems biology. Genome Inform;16:281–285.
4. Oprea T.I., Gottfries J. (2001) Chemography: the art of navigating in chemical space. J Comb Chem;3:157–166.
5. Aleksandrov A.D., Kolmogorov A.N., Laurent've M.A. (1986) Mathematics – Its Content, Methods, and Meaning, Volume 3 (Translated by K, Hirsch). Cambridge, Massachusetts: The MIT Press.
6. Todeschinin R., Consonni V. (2002) Handbook of Molecular Descriptors. Weinheim, Germany: Wiley-VCH Verlag GmbH.
7. Maggiora G.M., Shanmugasundaram V. (2004) Molecular Similarity Measures. In: Bajorath J., editor. Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery. Totowa, New Jersey: Humana Press;p. 1–50.
8. Carbó R., Calabuig B. (1990) Molecular similarity and quantum chemistry. In: Johnson M.A., Maggiora G.M., editors. Concepts and Applications of Molecular Similarity. New York: John Wiley & Sons;p. 147–171.
9. Jolliffe I.T. (2002) Principal Component Analysis, 2nd edn. New York: Springer.
10. Borg I., Groenen P. (1997) Modern Multidimensional Scaling – Theory and Applications. New York: Springer.
11. Rassokhin D.N., Lobanov V.S., Agrafiotis D.K. (2001) Nonlinear mapping of massive data sets by fuzzy clustering and neural networks. J Comput Chem;22:373–386.
12. Agrafiotis D.K. (2003) Stochastic proximity embedding. J Comp Chem;24:1215–1221.
13. Pearlman R.S., Smith K.M. (1998) Novel software tools for chemical diversity. Perspect Drug Discov Des;9:339–353.
14. Agrafiotis D.K., Rassokhin D.N. (2002) A fractal approach to selecting an appropriate bin size for cell-based diversity estimation. J Chem Inf Comput Sci;42:117–122.
15. Xue L., Bajorath J. (2002) Accurate portioning of compounds belonging to diverse activity classes. J Chem Inf Comput Sci;42:757–764.
16. Xue L., Stahura F.L., Bajorath J. (2004) Cell-Base Partitioning. In: Bajorath J., editor. Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery. New Jersey: Humana Press, Totowa;p. 279–289.
17. Rush J.A. (1999) Cell-based methods for sampling in high-dimensional spaces. In: Truhlar D.G., Howe W.J., Hopfinger A.J., Blaney J., Dammkoehler R.A., editors. Rational Drug Design. New York: Springer;p. 73–79.
18. Lam R.L.H., Welch W.J., Young S.S. (2005) Cell-based Binning Methods and Cell Coverage System for Molecule Selection. US Patent Number 6850876, Issued 1 February 2005. Assignee: Smithkline Beecham.
19. Maggiora G.M., Shanmugasundaram V., Lajiness M.S., Doman T.N., Schultz M.W. (2005) A practical strategy for directed compound acquisition. In: Oprea T.I., editor. Chemoinformatics in

Drug Discovery. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA;p. 317–332.

20. Downs G.M., Barnard J.M. (2002) Clustering methods and their uses in computational chemistry. Rev Comput Chem;18:1–40.
21. Holliday J.D., Rogers S.L., Willett P., Chen M., Madfouf M., Lawson K., Mullier G. (2004) Clustering files of chemical structures using the fuzzy *k*-means clustering method. J Chem Inf Comput Sci;44:894–902.
22. Böcker A., Dersen S., Schmidt E., Teckentrup A., Schneider G. (2005) A hierarchical clustering approach for large compound libraries. J Chem Inf Model;45:807–815.
23. Böcker A., Schneider G., Teckentrup A. (2006) NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. J Chem Inf Model;46:2220–2229.
24. Brewer M.L. (2007) Development of a spectral clustering method for the analysis of molecular data sets. J Chem Inf Model; 47:1727–1733.
25. Carhart R.E., Smith D.H., Venkataraghavan R. (1985) Atom pairs as molecular-features in structure-activity studies – definition and applications. J Chem Inf Comput Sci;25:64–73.
26. Willett P., Winterman V., Bawden D. (1986) Implementation of nearest-neighbor searching in an online chemical structure search system. J Chem Inf Comput Sci;26:36–41.
27. Hert J., Willett P., Wilton D.J., Acklin P., Azzaoui K., Jacoby E., Schuffenhauer A. (2005) Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. J Med Chem;48:7049–7054.
28. Dutta D., Guha R., Jurs P., Chen T. (2006) Scalable partitioning and exploration of chemical spaces using geometric hashing. J Chem Inf Model;46:321–333.
29. Guha R., Dutta D., Jurs P., Chen T. (2006) R-NN curves: an intuitive approach to outlier detection using a distance based method. J Chem Inf Model;46:1713–1722.
30. Shanmugasundaram V., Maggiora G.M., Lajiness M.S. (2005) Hit-directed nearest-neighbor searching. J Med Chem;48:240–248.
31. Agrafiotis D.K. (1997) On the use of information theory for assessing molecular diversity. J Chem Inf Comput Sci;37:576–580.
32. Yan A. (2006) Application of self-organizing maps in compound pattern recognition and combinatorial library design. Comb Chem High Throughput Screen;9:473–480.
33. Sheridan R.P., Kearsley S.K. (2002) Why do we need so many chemical similarity search methods? Drug Discov Today;7:903–911.
34. Ginn C.M.R., Willett P., Bradshaw J. (2000) Combination of molecular similarity measures using data fusion. Perspect Drug Discov Des;20:1–16.
35. Hert J., Willet P., Wilton D.J. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. J Chem Inf Comput Sci;44:1177–1185.
36. Willett P. (2005) Searching techniques for databases of two- and three-dimensional chemical structures. J Med Chem; 48:4183–4199.
37. Willett P. (2006) Similarity-based virtual screening using 2D fingerprints. Drug Discovery Today;11:1046–1053.
38. Hall D.L., McMullen S.A.H. (2004) Mathematical Techniques in Multisensor Data Fusion, 2nd edn. Norwood, Massachusetts: Artech House.
39. Klein L.A. (2004) Sensor and Data Fusion: A Tool for Information Assessment and Decision Making. Bellingham: SPIE Press.
40. Xue L., Stahura F.L., Godden J.W., Bajorath J. (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. J Chem Inf Comput Sci;41:746–753.
41. Schuffenhauer A. (2003) Similarity metrics for ligands reflecting the similarity of target proteins. J Chem Inf Comput Sci;43:391–405.
42. Hert J., Willett P., Wilton D.J., Acklin P., Azzaoui K., Jacoby E., Schuffenhauer A. (2006) New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. J Chem Inf Model;46:462–470.
43. Truchon J.-F., Bayly C.I. (2007) Evaluating virtual screening method: good and bad metrics for the ``early recognition'' problem. J Chem Inf Model;47:488–508.
44. Williams C. (2006) Reverse fingerprinting, similarity searching by group fusion and fingerprint bit importance. Mol Divers;10:311–332.
45. Tovar A., Eckert H., Bajorath J. (2007) Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. Chem Med Chem;2:208–217.
46. Wishart D.S., Knox C., Guo A.C., Shrivastava S., Hassanali M., Stothard P., Chang Z., Woolsey J. (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res;34:D668–D672.
47. Olah M., Mracec M., Ostopovici L., Rad R., Bora A., Hadaruga N., Olah I., Banda M., Simon Z., Mracec M., Oprea T.I. (2004) WOMBAT: world of molecular bioactivity. In: Oprea T.I., editor. Chemoinformatics in Drug Discovery. New York: Wiley-VCH;p. 221–239.
48. Willett P., Barnard J.P., Downs G.M. (1998) Chemical similarity searching. J Chem Inf Comput Sci;38:983–996.
49. Flower D.R. (1998) On the properties of bit-string based measures of chemical similarity. J Chem Inf Comput Sci;38:379–386.
50. Holliday J.D., Salim N., Whittle M., Willett P. (2003) Analysis and display of size dependence of chemical similarity coefficients. J Chem Inf Comput Sci;43:819–828.
51. Liu T., Lin Y., Wen X., Jorissen R.N., Gilson M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res;35:D198–D201.
52. Kearsley S.K., Sallamack S., Fluder E.M., Andose J.D., Mosley R.T., Sheridan R.P. (1996) Chemical similarity using physiochemical property descriptors. J Chem Inf Comput Sci;36:118–127.
53. Ginn C.M.R., Turner D.B., Willett P., Freguson A.M., Heritage T.W. (1997) Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. J Chem Inf Comput Sci;37:23–37.
54. Lootsma F.A. (1999) Multi-Criteria Decision Analysis via Ratio and Difference Judgement. Dordrecht, The Netherlands: Kluwer Academic Publishers.

55. Gillet V.J., Willett P., Fleming P.J., Green D.V.S. (2002) Design focused libraries using MoSELECT. J Mol Graph Model;20:491–498.
56. Agrafiotis D.K. (2002) Multiobjective optimization of combinatorial libraries. J Comput Aided Mol Des;16:335–356.
57. Gillet V.J. (2004) Applications of evolutionary computation in drug design. In: Johnston R.L., editor. Applications of Evolutionary Computation in Chemistry. New York: Springer;p. 133–152.
58. Soltanshahi F., Mansley T.E., Choi S., Clark R.D. (2006) Balancing focused combinatorial libraries based on multiple GPCR ligands. J Comput Aided Mol Des;20:529–538.
59. Lajiness M.S. (1997) Dissimilarity-based compound selection techniques. Perspect Drug Discov Des;7⁄8:65–84.
60. Irwin J.J., Shoichet B.K. (2005) Zinc – a free database of commercially available compounds for virtual screening. J Chem Inf Model;45:177–182.
61. Deb K. (2001) Multi-Objective Optimization using Evolutionary Optimization. New York: John Wiley & Sons.

Notes

[a] Developmental Therapeutics Program. Bethesda, MD: National Cancer Institute. Available at: http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html, accessed on 3 August 2007.

[b] Neurotransmitter.net. Available at: http://neurotransmitter.net/drug_reference.html, accessed on 3 August 2007.

[c] MACCS Structural Keys. San Leandro, CA: MDL Information Systems Inc. Available at: http://www.mdli.com, accessed on 21 March 2007.

[d] Molecular Operating Environment (MOE) (2005) Montreal, Quebec, Canada: Chemical Computing Group Inc. Available at: http://www.chemcomp.com, accessed on 21 March 2007.

[e] TIBCO Software, Inc., Spotfire US, Somerville, MA: TIBCO Software, Inc. Available at: http://www.spotfire.com, accessed on 17 July 2007.