

Diversity of Chemical Libraries

Dimitris K. Agrafiotis

3-Dimensional Pharmaceuticals, Inc., Exton, PA, USA

1	Introduction	742
2	Molecular Representation	742
3	Dimensionality Reduction	747
4	Diversity Metrics and Selection Algorithms	749
5	Visualization	754
6	Conclusions	759
7	Related Articles	760
8	References	760

Abbreviations

ACD = Available Chemicals Directory; CAS = Chemical Abstracts Service; CSD = Cambridge Structural Database; DEC = dynamically expanding context; FA = factor analysis; LVQ = learning vector quantization; MDS = multi-dimensional scaling; NNT = nearest-neighbor table; PCA = principal component analysis; QSAR = quantitative structure-activity relationships; SOM = self-organized map.

1 INTRODUCTION

Molecular diversity is the study of the structural, physicochemical, and biological heterogeneity of molecular collections. It has emerged at the dawn of a new era in experimental drug discovery, marked by the arrival of massively parallel synthesis and high-throughput screening. Traditionally, medicinal chemistry was based on a serial, systematic modification of chemical structure, using molecular similarity as a means to rationalize, organize, and prioritize experiments. It is based on the similar property principle,¹ that is the assumption that structurally similar compounds will exhibit similar physicochemical and biological properties. However, revolutionary advances in synthetic and screening technology have enabled the simultaneous synthesis and biological evaluation of large chemical libraries containing hundreds to tens of thousands of compounds. Molecular diversity has emerged as a result of this paradigm shift, and represents, in many ways, a generalization of the concept of molecular similarity from individuals to collections.

The study of molecular diversity is a new and incompletely understood field. It is closely related to molecular similarity, structure-activity correlation, and statistical series design, which are thoroughly reviewed in numerous texts.^{1,2} This article records our current awareness, and outlines the most important advances in terms of theory, methodology, and practice. It is divided into four sections that address issues related to molecular representation, dimensionality reduction, quantification and subset selection, and visualization. The reader is also referred to excellent reviews by Martin et al.,³ Blaney et al.,⁴ and Martin et al.⁵

2 MOLECULAR REPRESENTATION

2.1 Molecular Descriptors

2.1.1 Physicochemical Properties

The first attempts to quantify molecular diversity focused on physicochemical properties as a means of defining the diversity space. These properties can be calculated using standard molecular modeling and quantum mechanical packages, and include the dipole moment, HOMO and LUMO energies, heat of formation, total energy, ionization potential, number of filled orbitals, molecular weight, standard deviation of partial atomic charges and electron densities, octanol-water partition coefficient ($\log P$), molar refractivity, van der Waals volume and surface area, and many others. These properties have a long history in the field of structure-activity correlation. They capture steric, electronic, and lipophilic characteristics which are known to play a critical role in the transport and binding of a drug to its target. Molecular property descriptors have been used for diversity profiling by Willett,^{6,7} Martin et al.,⁸ Lewis et al.,⁹ Brown and Martin,^{10,11} and many others, and have been extensively reviewed by Kubinyi.¹²

2.1.2 Molecular Connectivity Indices

Molecular *connectivity* or topological indices (see **Topological Indices**) are numerical quantities based on certain invariants or characteristics of molecular graphs.^{13,14} They too have a long history in QSAR (see **Quantitative Structure-Activity Relationships in Drug Design**), and are designed to capture molecular properties such as size, ring structure, heteroatom content, branching, saturation, etc.

Topological indices can be divided into four main classes depending on their logical derivation: (1) indices based on the adjacency matrix, including the total adjacency index, the Zagreb group indices, the Randic connectivity index, the Platt index, the compatibility code, and the largest eigenvalue index; (2) indices based on the topological distance matrix, including the Wiener index, the polarity number, the distance sum, the Altenburg polynomial, the mean square distance, the Hosoya index, and the distance polynomial; (3) centric indices, including the generalized graph center; and (4) information-theoretic indices, including the Shannon index, the chromatic information index, the orbital information index, the topological information superindex, the electropy index, and the Merrifield and Simmons indices.^{13,14}

The use of topological indices for molecular diversity studies became popular owing to their minimal computational requirements and the widespread availability of Kier and Hall's Molconn-X suite.¹⁵ This program computes a wide range of topological indices based on molecular connectivity, shape, electrotopological state, topological equivalence, information content, and subgraph counts (number of rings, paths, clusters, etc.). Topological indices provide a convenient and inexpensive means of quantifying molecular structure, validated by years of experience in the field of structure-activity correlation.

2.1.3 Substructure Keys

Substructure keys encode molecular information in the form of binary arrays or bitmaps (see **Substructure Searching**). Each

element (or bit) in the array can take the values ‘true’ or ‘false’, and indicates the presence or absence of a specific structural feature or pattern in the target molecule. Substructure keys were originally designed for large-scale database searching, but have also proven effective in similarity applications.

To define a structural key, one defines the structural features of interest, assigns a bit to each one of these features, and generates the bitmap for each compound in the database. Typical features include the presence, absence, or minimum number of occurrences of a particular element (e.g., the presence of at least one, two, or three nitrogen atoms), unusual or important electronic configurations and atom types (e.g., doubly-bonded nitrogen or aromatic carbon), common functional groups such as alcohols, amines, etc., certain primitive and composite rings, and ‘disjunctions’ of unusual features that are rare enough not to be worth an individual bit, yet extremely important when they do occur (typically, these unusual features are assigned a common bit that is set if any one of the patterns is present in the target molecule). The generation of the keys is time-consuming since it requires a substructure search for each pattern represented in the bitmap and for each molecule in the database (a problem of quadratic or cubic time complexity). However, once the keys are generated, database searching involves Boolean operations between binary vectors, which are performed very rapidly by digital computers.

It should be noted that in most commercial database systems, the choice of structural features encoded in the substructure keys is aimed primarily at minimizing retrieval time. This is often application and database dependent. For instance, drug databases may encode functional groups of particular interest to medicinal chemists, while an organometallic database may contain features related to metal-containing functionalities. In any case, it turns out that these substructure keys contain sufficient information about the molecular structures to permit meaningful similarity comparisons. A number of similarity metrics have been proposed for binary descriptors (both substructure keys and molecular fingerprints; see below).⁶ The most frequently used ones are the normalized Hamming distance:

$$H = \frac{|\text{XOR}(x, y)|}{N} \quad (1)$$

which measures the number of bits that are different between x and y , the Tanimoto or Jaccard coefficient:

$$T = \frac{|\text{AND}(x, y)|}{|\text{IOR}(x, y)|} \quad (2)$$

which is a measure of the number of substructures shared by two molecules relative to the ones they could have in common, and the Dice coefficient:

$$D = \frac{2|\text{AND}(x, y)|}{|x| + |y|} \quad (3)$$

In the equations listed above, $\text{AND}(x, y)$ is the intersection of binary sets x and y (bits that are ‘on’ in both sets), $\text{IOR}(x, y)$ is the union or ‘inclusive or’ of x and y (bits that are ‘on’ in either x or y), XOR is the ‘exclusive or’ of x and y (bits that are ‘on’ in either x or y , but not both), $|x|$ is the number of bits that are ‘on’ in x , and N is the length of the binary sets measured in bits (a constant).

Another popular metric is the Euclidean distance which, in the case of binary sets, can be recast in the form:

$$E = \sqrt{N - |\text{XOR}(x, \text{NOT}(y))|} \quad (4)$$

where $\text{NOT}(x)$ denotes the binary complement of x . The expression $|\text{XOR}(x, \text{NOT}(y))|$ represents the number of bits that are identical in x and y (either 1s or 0s). The Euclidean distance is a good measure of similarity when the binary sets are relatively rich, and is mostly used in situations in which similarity is measured in a relative sense. Of all the indices mentioned above, Tanimoto is perhaps the most commonly used. As we will see below (Section 2.2), these simple substructure keys, when used with an appropriate clustering methodology, have been surprisingly successful in discriminating active from inactive compounds.

2.1.4 Hashed Fingerprints

One of the major shortcomings of substructure keys is the lack of generality. The information they encode and their effectiveness in database searching depends critically on the choice of features that go into their construction. For similarity studies, this is not necessarily a disadvantage, since it gives the designer the flexibility to select what he or she thinks are the most pertinent features for the application at hand.

Hashed fingerprints are a related class of descriptors that eliminate the arbitrary component from the encoding process. Just like structural keys, fingerprints were designed primarily for database applications, are binary in nature, and are derived directly from the connection table. Unlike structural keys, however, fingerprints do not depend on a predefined fragment dictionary to perform the bit assignment. Instead, every pattern in the molecule up to a predefined path length is systematically enumerated, and serves as input to a hashing algorithm that turns ‘on’ a small number of bits at pseudo-random positions in the bitmap. Because the number of possible patterns far exceeds the length of the fingerprint, bits are shared, so to speak, among a large but unknown number of patterns. For database applications this is not a problem, since every bit that is set in the pattern’s fingerprint will also be set in the molecule’s fingerprint which makes database screening deterministic and fast. This is not the case for similarity applications. However, although it is conceivable that two different molecules may have exactly the same fingerprint, the probability of this happening is extremely small for all but the simplest cases. Experience suggests that the similarity between two fingerprints is a good indicator of the similarity between the two structures. In addition, owing to their higher density and potential for overlapping patterns, fingerprints are more compact and have a higher discriminatory ability as the structures become more complex. A number of studies have shown that, at least for diversity analysis, fingerprints and substructure keys are just as effective (see Section 2.2).

Most of the descriptors discussed in this review require an enumerated structure, i.e., the full connection table of the reference molecule. For combinatorial libraries and for certain kinds of descriptors, this is not necessary. Downs and Barnard¹⁶ have recently presented an elegant method to compute molecular fingerprints based on the precursors, using techniques developed for Markush structure handling in chemical patents.

2.1.5 Atom Pairs and Topological Torsions

Atom pairs and topological torsions represent another solution to the problem of generality that plagues substructure keys.¹⁷ Atom pairs are patterns of the form $a_i - d - a_j$, where a_i and a_j are the types of atoms i and j , respectively, and d is their topological distance measured as the number of bonds along the shortest path connecting these atoms. A molecule with n atoms has $n(n-1)/2$ atom pairs, although some of them may not be unique. The topological torsion is of the form $a_i - a_j - a_k - a_m$, where i, j, k , and m are sequentially bonded atoms, and a_i is again the type of the i -th atom. In their original implementation, atom types included the atomic number, number of neighbors, and number of π electrons. Recently, however, these descriptors were expanded to include physicochemical¹⁸ and geometric¹⁹ features. Physicochemical atom pairs include binding properties, atomic $\log P$ contributions and partial atomic charges. In terms of binding properties, atom types are divided into seven classes: (1) anions, (2) cations, (3) neutral hydrogen bond donors, (4) neutral hydrogen bond acceptors, (5) polar atoms, (6) hydrophobic atoms, and (7) other. Hydrophobic and charge atom types are determined by assigning each atom to a pair of overlapping bins, and using the indices of the bins themselves as the actual atom types. Geometric atom pairs are similar to the regular atom pairs with the exception that the topological distance is replaced by the through-space distance of the corresponding atoms in some low-energy conformation of the target molecule.

The similarity between two structures is measured by:

$$s(i, j) = \frac{\sum_{k=1}^K \min(f_{ik}, f_{jk})}{0.5 \left[\sum_{k=1}^K f_{ik} + \sum_{k=1}^K f_{jk} \right]} \quad (5)$$

where f_{ik} is the count of the k -th descriptor in the i -th structure, and K is the union of all unique descriptors in i and j . This index ranges from 0 to 1, with 1 indicating complete identity and 0 indicating that the two structures have nothing in common.

Atom pairs and topological torsions have two desirable properties: (1) they are easy to compute; and (2) they can discriminate between closely related compounds. Geometric and topological atom pairs are equally effective in similarity searching, but the new generation of descriptors seem to perform worse than the original ones in their overall ability to discriminate biologically active from inactive compounds. As one would expect, which set does better than another varies greatly from probe to probe, and is very difficult to predict a priori.

2.1.6 Atom Layers

Atom layers were introduced by Martin et al.⁸ in order to capture the topological distribution of chemical features around a combinatorial core that are believed to play a critical role in receptor binding. The basic idea is that atoms that are closer to the attachment point contribute differently to binding than those that are more distant. Unlike most of the descriptors described so far, these properties apply only to combinatorial libraries and are therefore much more limited in scope. Atom

layers are constructed by summing a given property over all atoms in a side-chain at a given bond distance away from the attachment point. The properties considered by the Chiron group were the radius, and the acidic, basic, hydrogen bond donor, hydrogen bond acceptor, and aromatic character of the atom. The similarity between two substituents was computed by comparing the corresponding atom layer tables element by element, and dividing the sum of the minimum by the sum of the maximum values in each cell.

2.1.7 2D Autocorrelation Vectors

Autocorrelation is a technique that allows the compression of variable-length information into a fixed-length vector. The first application of this concept in the field of QSAR was presented by Moreau²⁰ who proposed an autocorrelation function to encode the topology of a molecular graph:

$$A(d) = \sum_{i,j} p_i, p_j \quad (6)$$

where p_i and p_j are an atomic property on atoms i and j , respectively, and d is the topological distance between the two atoms measured in bonds along the shortest path. Typically, only path lengths of size 2 to 8 are considered. Properties of interest include volume, electronegativity, hydrogen bonding character, and hydrophobicity. Autocorrelation vectors offer several advantages, in that they are compact, independent of the original atom numbering, and their length is independent of the size of the molecule.

In the original study, a separate autocorrelation vector was computed for each desired property, and the resulting set was reduced into a smaller number of variables using principal component analysis. Topological autocorrelation vectors were also used by Gasteiger²¹ as input to a Kohonen network which successfully separated dopamine from benzodiazepine receptor agonists, even when these compounds were buried among a large and diverse set of chemicals extracted from a commercial supplier catalog. Gasteiger²² has also extended the concept to three dimensions, by introducing a spatial autocorrelation vector based on properties measured on the molecular surface. These spatial autocorrelation vectors were used to model the activity of 31 steroids against the corticosteroid binding globulin, and the cytosolic Ah receptor activity of 78 polyhalogenated aromatic compounds. These descriptors were also found to be effective in describing the diversity of combinatorial libraries, also through the use of Kohonen networks (see Section 5.5).²³

2.1.8 B-cut Values

Searching for a universal definition of a chemical property space, Pearlman²⁴ found the solution to a class of descriptors inspired by Burden's work on molecular identification numbers.²⁵ Burden suggested that a molecular identification number could be constructed based on the two lowest eigenvalues of an $N \times N$ symmetric matrix representing the hydrogen-suppressed connection table of the molecule. Diagonal elements were replaced by the atomic numbers, while off-diagonal elements were assigned a value of 0.1 times the nominal bond type if the corresponding atoms were bonded, or 0.001 if the two atoms were not bonded, with an additional score of 0.01 being added to the (off-diagonal) terminal

nodes. Later, Rusinko and Lipkus applied Burden's concept for similarity searching using a 60 000-membered subset of the CAS (Chemical Abstracts Service) registry, and found that his approach compared well with results obtained from an established similarity searching procedure.

Pearlman extended this concept to multiple dimensions, by constructing three classes of matrices, using atomic charge, atomic polarizability, and hydrogen bond related properties across the diagonal, and a combination of interatomic distances, overlaps, computed bond orders, etc., off the diagonal. These properties were selected based on their importance in receptor–ligand recognition. A six-dimensional property space was then defined using the lowest and highest eigenvalues (B-Cut values) of a representative matrix from each of these three classes. Given the large number of combinations of diagonal and off-diagonal properties, the final choice was the one that produced the most uniform distribution of compounds in the chemical space, as determined by a χ^2 criterion. The properties themselves can be computed at multiple levels of theory, ranging in rigor and computational complexity. Pearlman reported that B-Cut values based solely on the connection table were proven satisfactory for most diversity profiling tasks.

The resulting six-dimensional space is small enough to permit diversity analysis based on partitioning (binning). This is described in greater detail below (Section 4.6).

2.1.9 3D Structural Keys

The use of three-dimensional information for diversity analysis appears obvious, if not necessary, if one considers the origins of biological specificity. Indeed, receptors and enzymes recognize shapes and electronic properties, rather than specific substructures or atom types. This idea, which can be traced back to Fischer's lock-and-key model and Ehrlich's pharmacophore hypothesis, has had a profound impact on modern drug design, experimental and computational alike.

3D structural keys are the extension of substructure keys in three dimensions. They are binary sets designed to function as screens for fast 3D database searching applications.^{26,27} Most of these systems involve two-center keys that record the distances or angles between two 'interesting' features such as particular atom types, centroids of aromatic rings, ring normals, and attachment points of functional groups. The key is computed by defining a set of atom types, and allocating a fixed number of bits to each unique pair. Each bit corresponds to a particular distance range for the two features. To generate the key, the features of interest are mapped onto the target molecule, and the distance of each pair is recorded onto the bitmap by turning on the corresponding bit.

Naturally, 3D structural keys require the availability of a three-dimensional structure of the target molecule. Earlier database systems relied on a single low-energy conformation, determined either crystallographically or computationally using a fast structure-generation software package.²⁸ These rigid screens were later modified to include flexible keys that take into account conformational flexibility. In the interest of efficiency, the conformational search procedure embedded in these systems is relatively crude and does not rule out highly strained conformations. One disadvantage of 3D structural keys is the poor representation of shape and chirality, and the limited number of features that can be effectively represented in a finite string. To address these limitations, several groups

have focused on a related class of descriptors known as 3D pharmacophore keys.

2.1.10 3D Pharmacophore Keys

The concept of a pharmacophore key was introduced by Sheridan and co-workers²⁹ as a means to account for the potential for intermolecular interactions in a 3D database search. Pharmacophore keys are 3D structural keys whose 'interesting' features include known macromolecular recognition sites. These sites include hydrogen bond donors, hydrogen bond acceptors, positively charged centers, aromatic ring centers, and hydrophobic centers. The pharmacophore itself is defined as a set of three or four centers forming a triangle or tetrahedron. To generate the key, the pharmacophores exhibited by a particular conformation or ensemble of conformations are mapped onto appropriate bits in the binary set. This process is illustrated in Figure 1.

Just like structural keys, pharmacophore keys can be readily extended to account for multiple conformations. Three-point pharmacophore keys also lend themselves to visualization in the form of a three-dimensional scatter plot (see Section 5.4). A number of people have followed up Sheridan's work, most notably the groups at Chemical Design,²⁷ Rhone-Poulenc,³⁰ and Abbott.¹⁰

One of the main uses of pharmacophore keys is to address questions related to molecular similarity and chemical diversity. Clearly, a diverse selection of compounds should exhibit a large number of pharmacophores, which should, in turn, be reflected in the union (or 'inclusive or') of their corresponding pharmacophore keys. Davies and Briant³¹ employed an iterative selection procedure that takes into account the flexibility of the compounds and the amount of overlap between their respective keys. The procedure is described in greater detail below (see Section 4.8).

2.1.11 Electronic Fields

The success of 3D-QSAR³² has prompted Cramer and co-workers³³ to develop descriptors based on the steric fields of single side-chain conformers 'topomerically' aligned around a common combinatorial core. This 'topomeric' alignment attempts to find a representative conformation for each side-chain attached to a particular variation site on a combinatorial template. The process starts with a low-energy conformation generated by a model-building routine, which is then fitted as a rigid body onto the template using least-squares minimization. The torsions of the rotatable bonds are then adjusted one by

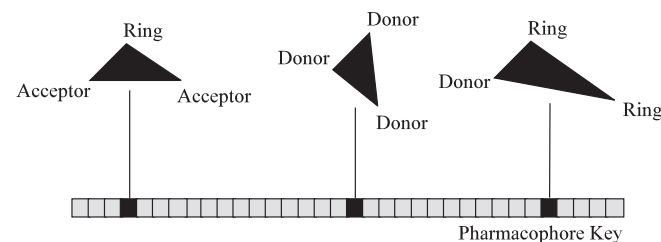


Figure 1 Three-point pharmacophore key generation. Each pharmacophore pattern present in the reference molecule is 'projected' onto a particular bit position, determined by the three 'atom' types and their mutual distances

one starting from the bond closest to the template, using a simple set of topological precedence rules. Once the alignment is complete, the steric field of the side-chain is computed using a CoMFA-like approach. These fields can then be used to compute a similarity index between two compounds using the root of the squares of the differences in steric field values summed over all lattice points in the CoMFA region, or some other equivalent distance function. The utility of these descriptors was demonstrated by classifying 736 commercially available thiols into 231 bioisosteric clusters which were consistent with results obtained from a clustering study based on 2D fingerprints and the Tanimoto similarity coefficient.

The 'topomeric' procedure described above attempts to circumvent the alignment problem that plagues most field-based QSAR techniques. While the assumption that template-based combinatorial compounds can be aligned with respect to their common core is certainly a reasonable one, it is not as clear how this technique could be used to compare multiple libraries or heterogeneous compound collections, or even libraries where the template itself is a variable site (a synthetic approach that is becoming quite common).

2.1.12 Affinity Fingerprints

While the vast majority of diversity analyses focus on the intrinsic structure of the individual compounds, there are some approaches that rely exclusively on the behavior of these compounds in their target environment. Such a biologically oriented approach was first introduced by researchers at Terrapin,³⁴ which built on the strengths of the company's fluorescent polarization high-throughput screening technology. In this approach, the molecular descriptor is the pattern of binding affinities (expressed in the form of IC₅₀ values) of a particular compound against a predefined set of protein targets. This vector, which is called the affinity fingerprint, can be used in much the same way as an ordinary multivariate descriptor to establish similarity relationships and perform diversity studies.

The choice of protein targets is of paramount importance if the results are to be statistically significant and of general utility. As a minimum requirement, the proteins must be able to recognize a wide range of organic compounds, which can be achieved by recording the binding information up to the limits of detection of the underlying biological assay. Moreover, the affinity patterns of these proteins must be uncorrelated and linearly separable. A systematic analysis of several hundred candidates resulted in a set of 18 proteins, which are now used routinely for screening new compound collections.

The similarity between affinity fingerprints provides an empirical way to assess the chemical variety of a chemical library. Kauvar³⁴ suggests that an estimate of the diversity of a given set of compounds can be determined based on the maximum separation and most frequently occurring distance between two affinity vectors in the collection. Although this is a rather crude and qualitative measure, the mathematical nature of the affinity fingerprint makes more robust diversity analyses (see Section 4) readily applicable. From a drug design perspective, affinity profiling has one important limitation: it cannot be used in a predictive mode for an unknown (untested) class of compounds. It is, however, conceivable that it may be possible to predict the biological profiles of known compounds against new protein targets through traditional regression techniques.

2.2 Descriptor Validation

As we pointed out in the introduction, the underlying assumption behind any diversity-based selection is that structurally similar compounds should exhibit similar physicochemical and biological properties. What is less clear is how one defines similarity and where the similarity cutoff lies; that is, how similar two compounds need to be for this statement to hold true.

Despite the lack of a universal and rigorous definition for molecular diversity, there seems to be general agreement that the choice of descriptors and similarity metrics should eventually be made on the basis of their ability to discriminate active from inactive compounds. The most comprehensive study specifically designed to address this issue was reported by Brown and Martin¹⁰ from Abbott Laboratories. What makes their contribution unique is the unprecedented size of the data set used in their analysis. In particular, four data sets were examined, containing in excess of 20 000 structures synthesized as part of past and on-going research projects at Abbott. This set included three different sets of compounds tested against monoamine oxidase and two other proprietary enzyme assays, and a collection of over 16 000 compounds that were tested over the years in the company's high-throughput screens.

Their aim was to identify which combination of descriptors and clustering methodologies were more effective in separating actives from inactives. The quality of the classification was measured by comparing the proportion of actives in clusters containing at least one active, to the proportion of actives in the data set as a whole, excluding, of course, any active singletons. In total, seven types of descriptors (MACCS, Unity and Daylight fingerprints, Unity 3D rigid and flexible descriptors, and two pharmacophore descriptors developed in-house), and four different clustering methodologies (Jarvis-Patrick, Ward, group-average, and Guenoche) were examined. They found that the 2D descriptors were considerably more effective, as was Ward's hierarchical agglomerative clustering algorithm. In a subsequent study, they also found that this trend is mirrored in the ability of these descriptors to predict other receptor-ligand recognition determinants such as hydrophobicity, dispersion, electrostatics, and steric and hydrogen bonding capability. These results were consistent with previous reports by the Sheffield group,⁷ and suggest that the existing 3D descriptors provide a rather poor representation of the receptor binding potential of a molecule, perhaps owing to the limited number of conformations and poor geometric descriptions involved. 3D descriptors are in general parsimonious, and do not provide the same density of sampling and careful selection of features that are typically involved in 3D QSAR analysis.

In a related study, Patterson et al.³⁵ reported an alternative method for validating descriptors based on the concept of a 'neighborhood radius'. Their objective was to develop a general method for identifying descriptors with a high discriminating power that would allow the detection of activity 'islands', i.e., regions in property space with a relatively high density of biologically active compounds. Their approach was to compare the differences in the descriptor values against the differences in the biological activities for a set of related compounds. If the descriptor is to be useful as a measure of similarity, the resulting plot should exhibit a characteristic trapezoidal distribution revealing a 'neighborhood behavior' for that descriptor.

The method was applied to 20 data sets, and 11 descriptors were ranked by their validity in series design. They concluded that 2D fingerprints and 3D CoMFA fields far outperformed physicochemical properties such as $\log P$ and molar refractivity, while topological descriptors such as connectivity indices, atom pairs and autocorrelation vectors fell in the middle of the spectrum. Interestingly enough, they also found that 2D fingerprints based on the whole molecule performed worse than those based on the side chains alone, which they attributed to a diluting effect by the template. Patterson's study considered one descriptor at a time, and did not account for the possibility of correlations between two or more descriptors.

3 DIMENSIONALITY REDUCTION

3.1 The Curse of Dimensionality

When dealing with multivariate molecular representations, one becomes quickly aware of a problem known as the curse of dimensionality. This term was first introduced by Bellman³⁶ to describe the complexity of combinatorial optimization over many dimensions, where the computational effort was found to scale exponentially with the dimensionality d . In statistics, this expression is used to describe the sparsity of data in higher dimensions. Indeed, our intuition based on two- and three-dimensional geometry is of limited use as we move to high-dimensional spaces. The following two examples illustrate the point. Consider, for instance, the fraction of the volume of a d -dimensional hypercube contained within the inscribed hypersphere:³⁷

$$f_d = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \quad (7)$$

For $d = 1, 2, 3, 4, 5, 6,$ and 7 , f_d is $1, 0.785, 0.524, 0.308, 0.164, 0.081,$ and 0.037 , respectively. It is clear that as d increases, the center of the hypercube becomes insignificant and its volume is concentrated near its corners. This apparent paradox has also been demonstrated by Wegman³⁸ by considering the hypervolume of a thin shell, i.e., the volume contained within two concentric hyperspheres, one with radius r and the other with a slightly smaller radius, $r - \varepsilon$. The fraction of the volume of the larger sphere contained within the two spheres is given by:

$$f_s = \frac{V_d(r) - V_d(r - \varepsilon)}{V_d(r)} = 1 - \left(1 - \frac{\varepsilon}{r}\right)^d \xrightarrow{d \rightarrow \infty} 1 \quad (8)$$

Thus, for higher dimensions, the volume of the hypersphere is concentrated mostly on its surface. According to equation (8), if the data is uniformly distributed, most of the samples will be found near the boundaries of the feature space. These two simple examples illustrate that the concept of 'neighborhoods' in higher dimensions is somewhat distorted; if the neighborhoods are 'local', they are virtually empty; if they are not empty, then they are not 'local'. This has important consequences in many statistical applications. There is, for example, a common ambition among many practitioners of combinatorial chemistry to synthesize a 'universal library' containing every possible pharmacophore. Some have gone as far as suggesting that this can be achieved with a mere 20 000 carefully selected compounds.³⁹ Whether this is possible depends on the definition of pharmacophore space and

the resolution of the sampling procedure. Many believe that pharmacological property space requires at least 20 unique dimensions.³ A hypercube of 20 dimensions has 2^{20} or 10^6 corners. Even for a coarse grid of only five values for each dimension, there are 10^{14} points that one needs to sample to fully explore the diversity space. While many of these points may be inaccessible due to chemically disallowed combinations, there is still an astronomical number of possibilities remaining that far exceeds the capacity of modern synthetic technology.

3.2 Dimensionality Reduction

For diversity profiling, high-dimensional representations pose a number of problems. The first, and perhaps most important, is the presence of substantial correlation between variables. Strictly speaking, correlation is not a problem caused by having many variables; it is quite possible to have complete correlation with only two or three variables, but the probability of this happening increases with the number of variables used. The importance of correlation depends on the application at hand, but in general redundant variables tend to exert undue influence in the analysis. Moreover, if the features are to be used for property prediction or classification (which is the natural next step of any diversity-based design), over-fitting can be a serious threat. The existence of a large number of variables can cause most regression and classification techniques to focus on the idiosyncrasies of the individual samples and lose sight of the broad picture that is essential for generalization beyond the training set. Finally, as the dimensionality of the space increases, the size of the computational effort needed to perform the analysis can be daunting even for today's most powerful computers.

Fortunately, most multivariate data in \mathfrak{R}^d are almost never d -dimensional. That is, the underlying structure of the data is almost always of dimensionality lower than d . In the interest of parsimony, and to simplify the analysis and representation of the data, it is often desirable to reduce the dimensionality of the space by eliminating dimensions that add very little to the overall picture. We must stress that none of the methods that will be discussed here guarantees to extract the most important features for the application at hand. There is always the possibility that some critical piece of information is left behind, buried under a pile of redundancies. Experience in many different application areas has shown that, in practice, this situation does not arise often.

This discussion will focus on two main techniques to perform the reduction: (1) principal component analysis; and (2) factor analysis. Both of these techniques attempt to find an appropriate low-dimensional representation of the covariance matrix. Other approaches such as multi-dimensional scaling, non-linear mapping, and Kohonen networks are reviewed briefly in this section, and discussed in greater detail in Section 5.

3.2.1 Principal Component Analysis

Principal component analysis (PCA) is a statistical technique with a long history in multivariate data analysis (see *Chemometrics: Multivariate View on Chemical Problems*).⁴⁰ PCA reduces a set of partially cross-correlated data into a smaller set of orthogonal variables (principal components)

without a significant loss in the contribution to variation. In effect, the method detects and combines descriptors which behave in a similar way into a new set of variables that are non-correlated, i.e., they are orthogonal.

Principal components are computed by diagonalizing the variance-covariance matrix. This is a square symmetric matrix whose diagonal elements represent the variances of each of the measured variables, and the off-diagonal elements represent their covariances. The elements of this matrix, m_{ij} , are computed using equation (9).

$$m_{ij} = m_{ji} = \frac{1}{N} \sum_{k=1}^N (x_{ki} - \mu_i)(x_{kj} - \mu_j) \quad (9)$$

where μ_i is the mean of variable i :

$$\mu_i = \frac{1}{N} \sum_{j=1}^N x_{ij} \quad (10)$$

and N is the number of observations in the data set. The eigenvectors of this matrix are the principal components, and the eigenvalues are their respective variances. Thus, the number of PCs is equal to the number of the original variables. However, if there is some redundancy in the data, it is usually sufficient to retain only the first few PCs that account for ‘most’ of the variance in the original data. This limit is arbitrary, and is usually determined by heuristic rules (typically, a threshold of 90 or 95% is used). Finally, the original data is transformed by using equation (11):

$$\mathbf{x}' = \mathbf{V}^T \mathbf{x} \quad (11)$$

where \mathbf{V}^T is the transpose of the eigenvector matrix, \mathbf{x} is the original multivariate sample, and \mathbf{x}' are the coordinates of that sample in the transformed space.

The main advantage of PCA is that it makes no assumptions about the probability distributions of the original variables. The elements of each eigenvector, which are called loadings, reflect the influence of the original variables in that eigenvector, and are used to establish natural associations between variables. PCA (as well as factor analysis) are sensitive to outliers, missing data, and poor correlations between variables due to poorly distributed variables. The method has been employed by a number of groups, including Martin et al.,⁸ Gibson et al.,⁴¹ and many others.

3.2.2 Factor Analysis

Factor analysis (FA) is a closely related technique that attempts to extract coherent subsets of variables that are relatively independent from one another (see *Chemometrics: Multivariate View on Chemical Problems*).⁴⁰ It is often the case in science that the variable we are interested in is not directly observable. However, it is often possible to measure other quantities that reflect the underlying variable of interest. Factor analysis is an attempt to explain the correlations between variables in the form of underlying factors, which are themselves not directly observable, and which are thought to be representative of the underlying process that has created these correlations.

Factors are linear combinations of original variables. They may be associated with two or more of these variables

(common factors) or with a single variable (unique factors). The specific association between the original variables and the derived factors is described in the form of loadings, which are derived from the magnitude of the eigenvalues of the covariance matrix. Factor loadings are inherently indeterminate. Rotation attempts to put these factors into a simple position, so that each variable is loaded highly on one factor, and all factor loadings are either large or near zero. A number of different rotation methods are available, including varimax, quartimax, and equimax. The varimax method maximizes the variance of the loadings, and is the most widely used.

On the surface, factor analysis and principal component analysis are very similar. Both rely on an eigenvalue analysis of the covariance matrix, and both use linear combinations of variables to explain a set of observations. However, in PCA the quantities of interest are the observed variables themselves; the combination of these variables is simply a means for simplifying their analysis and interpretation. Conversely, in factor analysis the observed variables are of little intrinsic value; what is of interest is the underlying factors.

Factor analysis has been used by Cummins et al.⁴² to reduce a set of 61 molecular properties to four factors, which were then used to compare the diversity of five chemical databases (see Section 4.6). It was also explored by Gibson et al.⁴¹ in a comparative study of 100 different heterocyclic aromatic systems, but they concluded that FA did not reduce the complexity of the analysis, and did not offer any significant advantages over PCA.

3.2.3 Multi-dimensional Scaling, Non-linear Mapping, and Kohonen Networks

Multi-dimensional scaling (MDS), non-linear mapping, and Kohonen networks (see *Neural Networks in Chemistry*) represent alternative techniques that can be used for dimensionality reduction. The first two were designed to reproduce coordinates from a distance or similarity matrix, while the latter features data abstraction by means of prototyping, achieved through a powerful self-organizing principle. In the first two techniques, the reduction is effected by reconstructing a low-dimensional coordinate set from a distance matrix computed from a higher-dimensional representation, while in the latter the original property vectors are onto a two-dimensional cell array arranged in a way that preserves the topology and density of the original data set. These reduced representations can subsequently be used for a variety of pattern recognition and classification tasks. Because these methods are ideally suited for visual inspection, they are discussed in greater detail in Section 5.

4 DIVERSITY METRICS AND SELECTION ALGORITHMS

4.1 Clustering

Because of its long tradition in molecular similarity applications, clustering was one of the first methods to be applied in diversity analysis (see *Chemometrics: Multivariate View on Chemical Problems*).⁶ In general terms, clustering is a

multi-variate analysis technique that seeks to organize information about variables in a set of relatively homogeneous groups called 'clusters'. These clusters must be internally homogeneous (members of a single cluster must be similar to each other) and externally heterogeneous (members of two different clusters must be dissimilar to each other). Contrary to other multivariate techniques, cluster analysis lacks a comprehensive body of statistical theory and is heuristic in nature.

There are four major steps in cluster analysis. First, an appropriate set of features must be selected, and scaled in a meaningful way. Second, a similarity matrix must be constructed, that records the distances between each pair of objects in the collection. Third, a decision must be made about the number and interpretation of the clusters. Finally, the cluster solution must be validated by visual or statistical means.

Clustering methodologies fall into two main categories based on the way in which the clusters are formed. Non-hierarchical clustering (also known as *k*-nearest-neighbor clustering) produces a single set of clusters based on some user-defined criteria. The most popular member of this family, particularly for diversity related tasks, is the Jarvis-Patrick algorithm. This method constructs the clusters by computing the *k* nearest neighbors of each object, and then clustering objects together if they are on each other's nearest neighbor lists and share some minimum number of nearest neighbors. The major advantage of this method is speed; its main disadvantage is its tendency to generate either too many singletons or too few very large clusters depending on the stringency of the clustering criteria. The second class is hierarchical clustering, whose output is represented in the form of a dendrogram or a tree. Hierarchical clustering may be top-down and employ logical division, or bottom-up and employ aggregation. In top-down clustering, the process starts with a single cluster which is recursively subdivided into smaller and smaller groups until each object is a member of its own cluster (singletons). Conversely, bottom-up approaches start with singletons and work their way to the top by combining clusters together to form larger collections. In general, cluster analysis is guided primarily by experience and heuristic rules, and requires that the user makes certain decisions which may have a profound influence on the results of the classification.

The most complete study of similarity clustering in chemical systems was presented by Willett⁷ who carried out a systematic comparison of four different clustering methodologies, including the Ward and group-average hierarchical agglomerative methods, the minimum diameter polythetic hierarchical divisive method, and the Jarvis-Patrick nearest neighbor algorithm. The methods were tested on a set of 5982 compounds characterized by 13 molecular properties, and the results were evaluated by means of simulated property prediction experiments. These experiments suggested that the Jarvis-Patrick algorithm is not an appropriate choice for clustering molecular property data, and that any of the first three methods should be preferred. A subsequent study by Brown and Martin has confirmed these findings.¹⁰

In order to design a diverse series, the user typically selects a representative from each cluster (usually the centroid), and then optionally examines this resulting subset for possible multi-collinearities. If the selection is non-orthogonal, suspect compounds are replaced with other members of the same cluster, and the new solution is re-evaluated. This cycle continues until a quasi-orthogonal set is identified.

4.2 Maximin

Among all the selection algorithms presented in this review, maximin stands out for its conceptual simplicity and ease of implementation, though not necessarily for its efficiency. Maximin is a greedy algorithm that attempts to maximize the minimum intermolecular dissimilarity among a set of compounds. The process starts with a randomly chosen compound, and builds up the selection by adding one compound at a time. During each iteration, the algorithm evaluates all the remaining candidates (that is, every compound that is not already a member of the selection), and selects the one whose distance from its nearest neighbor in the existing set is maximal. This procedure scales to the square of the number of compounds being considered, and becomes impractical for larger designs. An alternative implementation was presented by Hassan⁴³ and Agrafiotis,⁴⁴ who used the maximin criterion itself as a diversity function that was optimized using a simulated annealing approach. The first application of maximin was reported by Lajiness,⁴⁵ but many people have followed suit, including Polinsky,⁴⁶ Chapman⁴⁷ and others.

4.3 Stepwise Elimination

Many of the algorithms discussed in this section construct the solution from the bottom up, that is starting from the null set and incrementally augmenting this set until the desirable number is reached. Taylor⁴⁸ reported an alternative technique which works in the opposite direction. The method starts with the full set, and eliminates one compound at a time based on the maximum similarity principle. In particular, the $N \times N$ similarity matrix is scanned to identify the largest element, and one of the compounds associated with that element is selected at random and eliminated. The process continues until a single compound is left in the set. This algorithm sorts the compounds in reverse order of dissimilarity, placing the most diverse molecules at the top of the list.

4.4 Cluster Sampling

Another technique developed by Taylor⁴⁸ was cluster sampling. This is a nearest-neighbor approach that, unlike conventional cluster analysis, never explicitly partitions the space into a set of clusters. The method starts by constructing a nearest-neighbor list for every compound in the database, using a similarity threshold of 0.8 (minimum similarity necessary for a pair of compounds to be considered neighbors). The nearest-neighbor lists of all the compounds are combined to form the nearest-neighbor table (NNT) of the database. Having constructed the NNT, the selection algorithm proceeds as follows: the first molecule to be extracted is the one that occurs most often in the NNT, since this would tend to be at the center of the most densely populated region in the property space. All nearest neighbors of the selected molecule are then marked as 'held', which means that they become unavailable for selection. The next compound selected is the one that occurs most often in the NNT and is not marked as held. This is typically the center of the second most densely populated region in the space. This compound's neighbors are then held, and the process continues until all compounds are either chosen or held. Both cluster sampling and stepwise elimination are intuitive and robust procedures, but scale to the square of *N*, which makes them impractical for large data sets.

4.5 Experimental Design

In what is still considered by many the most complete work in the field, Martin and co-workers at Chiron⁸ reported a selection technique based on the principles of statistical experimental design (see *Experimental Design*). Their objective was to provide a rational method for selecting a representative set of amines and carboxylic acids to be used as side-chains and capping groups in *N*-substituted glycine peptoid combinatorial libraries.

It was clear that the set of descriptors used to define the diversity space would play a critical role in the outcome of the design. To ensure that the design would capture biologically relevant information, the group employed a wide range of molecular descriptors that were classified in four broad classes: (1) lipophilicity; (2) shape and branching; (3) chemical functionality; and (4) receptor binding. Lipophilicity was described by means of octanol–water partition coefficients, calculated by $C \log P$, HINT, or LOGKOW. Shape, flexibility, branching, and ring structure were described by a set of 81 topological indices, including the molecular weight, the number of elements, heavy atoms and bonds, 70 connectivity indices, and seven shape descriptors (see Section 2.1.2). Using principal component analysis, these 81 descriptors were subsequently reduced to five latent variables which captured 86% and 87% of the variance of the amine and carboxylate data sets, respectively. The chemical functionality descriptors were computed indirectly, by first constructing a pairwise distance matrix using Daylight fingerprints and the Tanimoto similarity coefficient, followed by multi-dimensional scaling (MDS, see Section 5.6) to reduce these fingerprints to a small number of continuous variables (5 to 7) that reproduced all the original dissimilarities with an error of around 10%. Finally, the receptor binding potential of each side-chain was described by atom layers emanating from the peptoid backbone and describing hydrophobic, hydrogen bonding, and polar functionalities in the form of an autocorrelation-like data table. These tables were then used to compute pairwise similarities, which were subsequently reduced to five continuous variables using classical MDS. These descriptors were combined to form a set consisting of $\log P$, five shape and branching PCs (or seven in the case of the carboxylates), five chemical functionality MDS dimensions, and five receptor recognition MDS dimensions. Within each set, the descriptors were rescaled to zero mean and unit variance for the first component, assigning, in effect, lesser weight to the remaining PCs and MDS dimensions.

Having defined a suitable property space, the actual selection was carried out using a statistical technique known as D-optimal design. D-optimal design takes as input a design matrix, \mathbf{X} , and selects a subset of points from a larger pool of candidates that maximize the determinant of the ‘information matrix’, $\mathbf{X}^T\mathbf{X}$. The rows of \mathbf{X} represent the monomers or compounds, and the columns represent the original features, or higher-order terms such as their squares, cubes, or cross terms. This, in effect, minimizes the determinant of the inverse, which is the variance of the parameter estimates for the model that is encoded by the design matrix. The actual algorithm builds up the design in a stepwise manner, starting with the null set (or some pre-selected set of compounds), and gradually augmenting that set by including the monomer that best complements the existing solution, until the maximum number is reached.

The objective of D-optimal design is to identify points that are not only spread out in property space, but are also orthogonal. This is achieved by maximizing the volume in covariance space; the determinant is large when the variances are large (i.e., the points are spread out) and when the co-variances are small (i.e., they are orthogonal). D-optimality is, in fact, one of many different criteria that can be employed to achieve this task. Other criteria include A-, G-, and I-optimality which are also model-based, and S- and U-optimality which are distance-based and are designed to fill up space and maximize spread.

While the method is statistically sound, it is not quite clear just how much spread is sacrificed to achieve a higher rank. Hassan et al.⁴³ have recently shown that the loss may in fact be quite substantial. Their idea was to cast the D-optimal criterion in the form of an objective function that measured the diversity of a given set of compounds, and then maximize this function using a Monte Carlo technique to identify the most diverse set. Their approach, which is discussed in greater detail below (see Section 4.7), showed that the D-optimal criterion favors the extremes of the feature space, and tends to ignore the central region. This is particularly true when the number of points selected far exceeds the dimensionality of the space. However, Hassan’s analysis was based on a linear model (i.e., the design matrix did not include any higher-order terms) which may partially account for this redundancy (when the model contains fewer terms than the number of observations, duplicates introduce an estimate of uncertainty in the response surface). It is possible that the inclusion of higher-order terms will partially compensate for this redundancy, but this remains to be demonstrated. It is also not clear whether the method can handle large data sets, and become a more general experimental design tool for drug discovery.

Finally, we should point out that Martin’s work is not the first application of experimental design for compound selection. These principles have been practiced for many years in statistical series design and structure–activity correlation. Marsili and Saller⁴⁹ in 1993, reported an almost identical approach for selecting multivariate synthetic analogs, using the determinant of the covariance matrix and a fractional factorial design procedure, implemented in the program *Analogs*. At that time, however, combinatorial chemistry was still at its infancy, and that work was largely ignored.

4.6 Partitioning Techniques

The quadratic time complexity of most diversity algorithms has led a number of groups to investigate alternative partitioning techniques. These techniques partition the property space into a set of multi-dimensional cells, by dividing each axis into a finite number of equally sized bins. Assigning a compound to a particular cell is a direct and extremely fast indexing operation. The method is flexible and intuitive, and is ideally suited to a variety of diversity-related tasks. Identifying and filling diversity voids is straightforward, as is selecting diverse subsets from a larger collection, and comparing the redundancy, complementarity, and diversity content of multiple collections. However, in order to truncate the combinatorially explosive enumeration of discrete cells, the method requires a low-dimensional representation of chemical space. This problem was addressed differently by different groups.

Pearlman's²⁴ solution was to construct a six-dimensional space using the highest and lowest eigenvalues (B-Cut values) of three atom association matrices which encoded charge, atomic polarizability, and hydrogen bonding information (see Section 2.1.8). The main problems with this approach are the lack of a physical interpretation and rather unclear significance of the extreme eigenvalues (Burden contests that the smallest eigenvalue contains contributions from every atom, and therefore reflects the topology of the entire molecule). Also, there is plenty of evidence suggesting that an adequate description of chemical space requires substantially more than six dimensions.⁸ But perhaps the greatest shortcoming of Pearlman's approach is the lack of published results supporting his work.

Cummins et al.⁴² used a similar approach to compare the diversity of five chemical databases containing in excess of 300 000 structures. These included two drug databases (CMC and MDDR), two databases of commercially available compounds (ACD (Available Chemicals Directory) and SPECS), and a collection of proprietary compounds synthesized as part of the research efforts at the Wellcome Foundation, known as the Wellcome Registry. The initial set of descriptors included a standard set of topological indices and an estimate of the free energy of solvation. Highly correlated or discrete features were eliminated, and the resulting set of 61 properties was subjected to factor analysis which identified four factors that explained 90% of the variance in the data. The diversity of each data set was then computed as the fraction of the total volume occupied by that set, using a Riemann-style approach. A trimming procedure was also employed to eliminate outliers and focus the analysis on the more densely populated areas of

the feature space. The resulting density functions for two of these factors are shown in Figure 2.

4.7 Stochastic Techniques

Most of the algorithms described so far are greedy in nature. That is, the selection proceeds in a stepwise manner by making decisions that make the most sense at the time, but do not actually guarantee that the optimal solution will be found. Selecting subsets of compounds from larger collections is a combinatorial problem of enormous proportions. In its simplest form, this is the infamous n -choose- k problem: given an n -membered collection and a number k , find the k most diverse compounds in that population. The number of different k -membered subsets of an n -membered set is given by the binomial:

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \quad (12)$$

This problem is NP-complete, and the cardinality of that space is enormous even for the most conservative cases encountered in combinatorial design. This has prompted Agrafiotis^{44,50} and Hassan et al.⁴³ independently to propose an alternative solution based on Monte Carlo sampling. Their idea was to define an objective function that measures the diversity of any conceivable subset of compounds, and then use simulated annealing to identify the optimal set.

Simulated annealing is a global, multivariate optimization technique based on the Metropolis Monte Carlo search algorithm. The method starts from an initial random state, and walks through the state space associated with the problem of

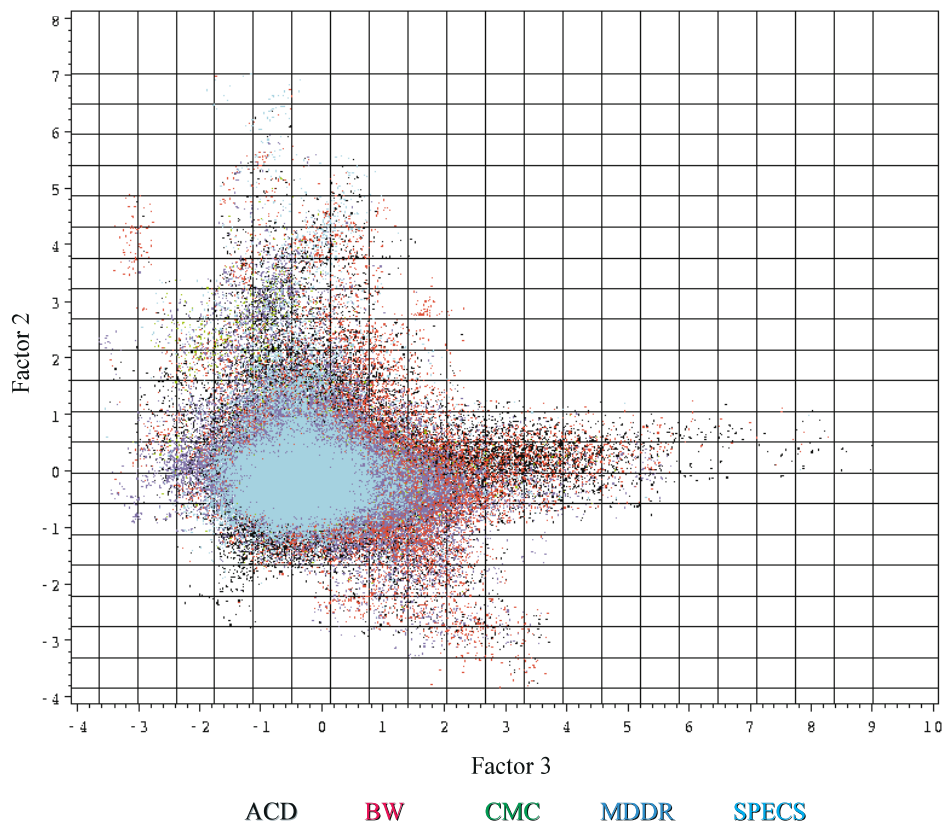


Figure 2 A comparison of the property distributions of five chemical databases using factor analysis

interest by generating a series of small, stochastic steps. An objective function maps each state into a value in \mathfrak{R} that measures its fitness. In the problem at hand, a state is a unique k -membered subset of compounds from the n -membered set, its fitness is the diversity associated with that set, and the step is a small change in the composition of that set (usually of the order of 1–10% of the points comprising the set). While downhill transitions are always accepted, uphill transitions are accepted with a probability that is inversely proportional to the energy (fitness) difference between the two states (usually $\exp(-\Delta E/k_B T)$).

Hassan used the maximin, power-sum, and product functions, while Agrafiotis' original implementation was based on maximin and a volumetric diversity measure of his own device. The convergent nature of this algorithm is shown in Figure 3 for the case of a simple 2D uniform data set:

Although their methods were similar, the motivations of the two groups appear to be rather different. Hassan et al. wanted to compare the performance of different diversity metrics, while Agrafiotis was interested in a much more general method that could encode any desirable selection criterion (similarity, predicted activity, cost and availability of starting materials, reaction block design, etc.) and would allow complex multi-objective selections in advanced decision support systems. His work was the first detailed report on an ambitious project known as DirectedDiversity[®],^{51,52} which integrates combinatorial, structural, and computational chemistry under a unifying data management system. He later applied this algorithm to study other diversity metrics with surprising results^{53,54} (see Sections 4.11 and 4.12), and has also reported evolutionary and genetic variants of this general sampling approach.

4.8 Boolean Logic

Diversity analysis is greatly simplified if the chemical space is of binary nature. Unlike continuous properties, binary descriptors such as structural keys and hashed fingerprints can be compared using fast binary operations to render quick estimates of molecular similarity, diversity, and complementarity. For example, the similarity between two compounds can be computed as a function of the intersection of their respective descriptors (logical AND), whereas diversity can be estimated based on their union (logical OR). This can be exploited in a number of different ways, elegant examples of which can be found in the work of Shemetulskis,⁵⁵ Pickett,³⁰ and Davies.³¹

4.9 Conformational Sampling

Chapman⁴⁷ has recently reported a rather ambitious substituent-based algorithm that makes explicit use of multiple low-energy conformations. The approach is based on a similarity index that measures how well two different conformers can be superimposed in terms of their steric and polar characteristics. The conformers may belong to the same or different molecules, and are aligned using an analytical continuous function and a gradient minimization procedure. The objective of this algorithm is to find an orientation that maximizes the overlap of steric bulk and polar functionalities. Once the conformers are aligned, the steric component of similarity is computed by summing the distances of each atom in each conformer from the nearest atom in the other conformer. Similarly, the polar component is computed by summing the distances of each partially or formally charged group of each conformer from the nearest group of the same sign of the other conformer.

To compute the diversity of a set of compounds, each molecule is subjected to a systematic conformational search, and the lowest-energy conformations are recorded. Each conformer of each compound is then compared to every other conformer of every compound, and the diversity of the system is computed using equation (13):

$$D(M) = \sum_{m \in M} \left[\left(\sum_{c \in C(m)} \min_{c' \in C} (d(c, c')) \right) - T\Delta S(m) \right] \quad (13)$$

where M is the set of all compounds, $C(m)$ is the set of all conformers of compound m , and C the set of all conformers of all compounds. $T\Delta S(m)$ is an entropic term that penalizes highly flexible compounds, and is a function of the number of rotatable bonds. The selection of the most diverse set is carried out using a greedy algorithm akin to maximin. The system builds up the solution incrementally, starting with the null set, and adding one compound at a time until the required number is reached. Each compound is chosen by evaluating the diversity that would be added by each remaining candidate to the existing set, and selecting the one that contributes the most.

The merits of this approach were assessed in two ways. First, by computing the similarity matrix of the naturally occurring amino acids using the measure described above, and, second, by testing the diversity selection procedure on a set of 1371 commercially available carboxylic acids. In the case of the amino acids, the results were intuitive and suggested that

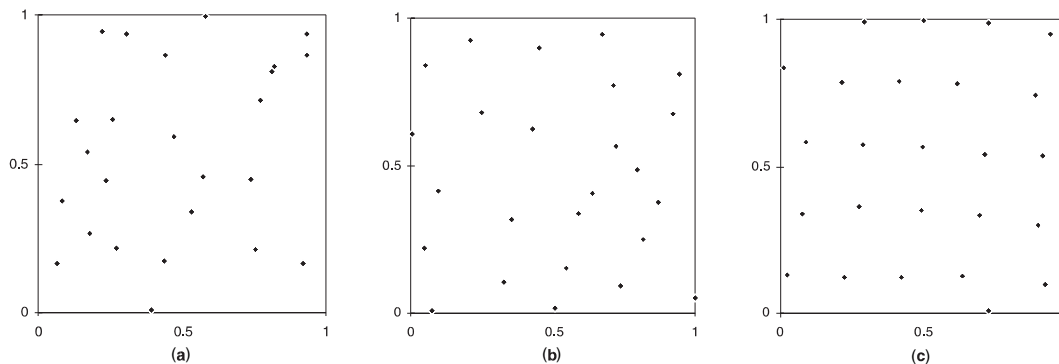


Figure 3 Selection of 25 diverse points from a uniform 2D data set using Agrafiotis' simulated annealing approach: (a) initial state; (b) best state found after first cycle; (c) best state found after 31st cycle

the measure of similarity is indeed a reasonable one. In the case of the carboxylates, the selection compared favorably to random controls, and identified reagents that were quite diverse in terms of shape, size, and functionality.

To say that this method is computationally intensive is probably an understatement. As a result, it can be applied only to small collections, and is probably best suited for screening reagents for combinatorial libraries. Even then, some topological and heuristic pruning is necessary to truncate the number of compounds and reduce the complexity of the task. While the method is intellectually robust and intuitive, it remains to be seen whether it offers any significant advantages over existing techniques to justify the extra effort required.

4.10 Vector Analysis

A couple of groups have reported diversity methods based on an analysis of the spatial relationships of intramolecular functionalities. Boyd⁵⁶ reported a method, called HookSpace, that measures diversity based on the spatial distribution of distances between user-defined functional groups. In particular, each pair of functional groups in a given compound was aligned on the xy plane so that one of the groups was placed along the x axis with the head atom at the origin, and the other was positioned on the xy plane, with the head-to-tail vector pointing in the positive z direction. Once the alignment was complete, the position of the head atom of the second group on the xy plane was recorded. This process was repeated for every pair of functional groups in each structure, and for every structure in the database. The xy plane was then partitioned into a finite number of cells, and each cell recorded either the total number of functional groups, or the number of different functional groups at that position. This permitted diversity measurements by computing the percentage of non-vacant cells on the xy plane, similar to the method described by Pearlman. The authors used this approach to compare the structural diversity of the Available Chemicals Directory (ACD), the Cambridge Structural Database (CSD), and a benzodiazepam combinatorial library, using a theoretical reference space. They concluded that the ACD covered 85% of that space, whereas the CSD and the benzodiazepam library covered only 34% and 13% of the space, respectively. It is quite likely, however, that this difference reflects the different origins of the three-dimensional structures of these compounds (computed versus experimental), rather than the intrinsic functional and geometric diversity of the two databases.

In a related approach, Bartlett⁵⁷ presented a system that compared the diversity of different combinatorial templates using the angles between the bond vectors connecting the core to the substituent. The method followed the spirit of the Caveat approach, and the results were presented in a visual form.

4.11 Information Theory

Lin⁵⁸ has proposed an alternative method for quantifying diversity based on Shannon's information theory. His method is based on the recognition that a diversity design attempts to maximize the information content of the resulting selection, and that this could be quantified using Shannon's classical entropy equation. The basic concept in Lin's approach is that every collection of compounds represents a finite number of distinguishable species, and that the 'distinguishability' of

these species can be described as a function of their mutual dissimilarity. The more distinguishable the species, the greater their information content. To cast this idea in the form of an equation, Lin proposed that the diversity (or information content) of a system be described as a function of the entropy:

$$I = S_{\max} - S \quad (14)$$

where

$$S = - \sum_{i=1}^N \sum_{j=1}^N p_{ij} \ln p_{ij} \quad (15)$$

N is the total number of compounds, p_{ij} is the probability of finding the i -th individual in the j -th species (given by some function of their dissimilarity), and S_{\max} is the maximum entropy of the system.

While the use of information theory seems like a natural choice, the actual implementation suffers from a number of disadvantages. In a recent article,⁵³ we reported that a strict application of this approach produced extremely unbalanced designs, and clustered points at maximum separation along the diagonal of the feature space. We believe that this is due to the use of the wrong type of 'information', and to the implicit assumption that ideal designs should be equiprobable (i.e., that the pairwise intermolecular dissimilarities should be as uniform as possible). In a personal communication, Lin suggested that our results could be an artifact of the similarity measure used in our study, but a detailed response has yet to appear in print. As of this writing, the debate is still open.

4.12 Cosine Coefficient

To circumvent the quadratic time complexity that plague many diversity algorithms, Willett and co-workers⁵⁹ proposed an alternative approach based on the cosine coefficient of similarity. In this approach, the diversity of a set of compounds, A , be defined by the mean pairwise intermolecular dissimilarity:

$$D(A) = 1 - \frac{\sum_{i=1}^N \sum_{j=1}^N \sigma(i, j)}{N^2} \quad (16)$$

where $\sigma(i, j)$ is the similarity between two compounds, i and j , in A , and N is the total number of compounds in A . Like many indices of this kind, in order to evaluate $D(A)$ one needs to compute the similarity matrix $\sigma(i, j)$, which scales to the square of the number of compounds in the data set. However, the authors showed that if the cosine coefficient is employed to compute the pairwise similarities, Equation (16) can be transformed into a functional form that can be evaluated in linear time:

$$D(A) = 1 - \frac{\mathbf{a}_c \cdot \mathbf{a}_c}{N^2} \quad (17)$$

where

$$\mathbf{a}_c = \sum_{i=1}^N w(i) \mathbf{m}(i) \quad (18)$$

The ‘weights’ $w(i)$ are given by:

$$w(i) = \frac{1}{\sqrt{\sum_{k=1}^K m(i, k)}} \quad (19)$$

where $\mathbf{a}_c \cdot \mathbf{a}_c$ is the dot product of the vector \mathbf{a}_c with itself, and $m(i, k)$ is the k -th property of the i -th compound.

The cosine coefficient is applicable to any situation in which the compounds can be represented in a vectorial form, e.g., by a set of topological indices or computed molecular properties. It has been used extensively in information retrieval systems, and is defined as the cosine of the angle formed by two molecular property vectors:

$$\sigma(i, j) = \cos(i, j) = \frac{\sum_{k=1}^K m(i, k)m(j, k)}{\sqrt{\sum_{k=1}^K m(i, k)^2 \sum_{k=1}^K m(j, k)^2}} \quad (20)$$

Unfortunately, this dramatic performance improvement comes at a significant price. Using simple trigonometric arguments and the stochastic approach outlined above, we showed that the method has a general tendency to sample the principal axes of the feature space and produce heavy redundancies.⁵⁴ This is due to the nature of the cosine coefficient which looks only at the angles between the property vectors and ignores their distances, and, more importantly, to the use of a simple dissimilarity summation function for measuring diversity.

5 VISUALIZATION

The most difficult challenge in data analysis is to be able to represent whatever complexities might be intrinsic to the data in a simple and intuitive form. Visualization is an important component of diversity analysis, and has attracted the interest of many groups.

5.1 Histograms and Multivariate Scatter Diagrams

The presentation of multivariate data is often accomplished in a tabular form, particularly with small data sets with named or labeled objects. Histograms provide a convenient means of analyzing one variable at a time, but this type of variable-by-variable examination of multivariate data can be overwhelming, and does not reveal any relationships between variables. 2D scatter diagrams are more effective in this respect, but they also present problems if the number of points is substantial, and the dimensionality of the space is relatively high. In this case, one needs to examine $d(d-1)/2$ pairwise plots, which can be overwhelming if d is larger than 5 or 6. A technique known as brushing allows the analyst to highlight a subset of points in a collection of graphs using a pointing device. One way in which this can be achieved is by linked graphs; a particular subset of points is highlighted on one graph using a distinct color or synchronous blinking, and that characteristic is inherited in the linked graphs, including other scatterplots, histograms, and regression plots. Even this technique, however, has limitations. If there are more than a

few variables, the eye cannot follow many of the dynamic changes in the pattern of points during brushing.

5.2 Chernoff Faces, Andrews’ Curves, Star Diagrams, and Parallel Coordinate Plots

The methods described in this section have not actually been applied in diversity studies. They are, however, quite common in multivariate analysis and are presented here for the sake of completeness.

Chernoff faces are simple glyphs that associate variables with facial features, such as the size and shape of the mouth, eyes, nose, and facial outline.⁶⁰ The motivation was the recognition that humans are extremely capable of discriminating faces, and that traditional visualization methods seemed to be less valuable in producing an emotional response. While this may be true for most multivariate data, it is certainly not the case with chemical structure diagrams.

Andrews⁶¹ proposed representing each point in R^d as a curve, $s(t)$, computed by a Fourier series of the original variables:

$$s(t|x_1, x_2, \dots, x_d) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \dots \quad (21)$$

This is an intriguing idea, as it allows each original multivariate point to be recovered from the curve. Since the Andrews’ curve is dominated by the low-frequency terms, some arbitrary decision must be made regarding the relative importance of the original variables.

A simpler and perhaps more relevant representation is the star.⁶² The original axes are drawn as spokes on a wheel, and the coordinate data values are plotted on these axes, and connected by line segments. A more aesthetically pleasing version was presented by Martin et al. in the form of a flower plot (see below).

Another intriguing approach is Inselberg’s parallel coordinate plot.⁶³ The plot consists of d evenly spaced parallel axes, representing the original variables. Each multivariate sample is represented as a piecewise linear curve connecting d -points on these parallel axes. The disadvantage is that points which share a common value in any of the dimensions cannot be distinguished without use of color.

All methods described above are most valuable with small data sets where the individual points are clearly identifiable. When applied to large data sets like those encountered in combinatorial designs, these methods have a tendency to produce ‘too much ink’ and confusion. One could always plot a subset of the data (a process known as thinning), but this usually compromises the interpretation.

5.3 Flower Plots

Martin et al.⁸ introduced flower plots as a means simultaneously to display all 16 molecular properties associated with a given side chain in a peptoid library. Flower plots are colorful variants of the star diagrams mentioned above. They are circular bar graphs, with one ‘petal’ for each molecular property or descriptor. Positive petals point outward, while negative ones point towards the center of the graph. The size of the petal is proportional to the value of that property, and the central sphere is color-coded according to some additional

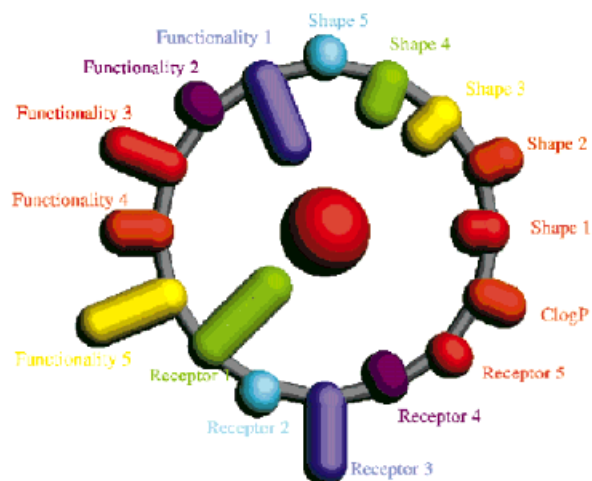


Figure 4 Flower plot of tyramine. There is one petal for each of the five chemical functionality descriptors, five shape descriptors, five receptor recognition descriptors, and the computed $\log P$

property such as a biological activity or similarity to a reference compound. Figure 4 shows the flower plot of tyramine,

described by means of five chemical functionality descriptors, five shape descriptors, five atom-layer receptor recognition descriptors, and the computed $\log P$.

Flower plots are particularly useful in assessing the distribution of properties across a collection of compounds. Figure 5 shows the structures and associated flower plots of 18 side-chains from a biased *N*-substituted glycine peptoid combinatorial library. The structures on the first row are tyramine itself and its closest analogs, while those in the lower rows are side-chains chosen using a D-optimal experimental design procedure from a pool of 721 amines (see Section 4.5). Clearly, the descriptors capture the structural similarity of the side-chains, and this is nicely reflected in the flower plots. However, these graphs become impractical for larger data sets, and it is not very clear (at least not to the author) that they aid more in the perception of similarity than the structural diagrams themselves.

5.4 Pharmacophore Plots

'Pharmacophore plots' are three-dimensional scatter diagrams that represent the three-point pharmacophores exhibited by a particular structure. The axes represent the distances

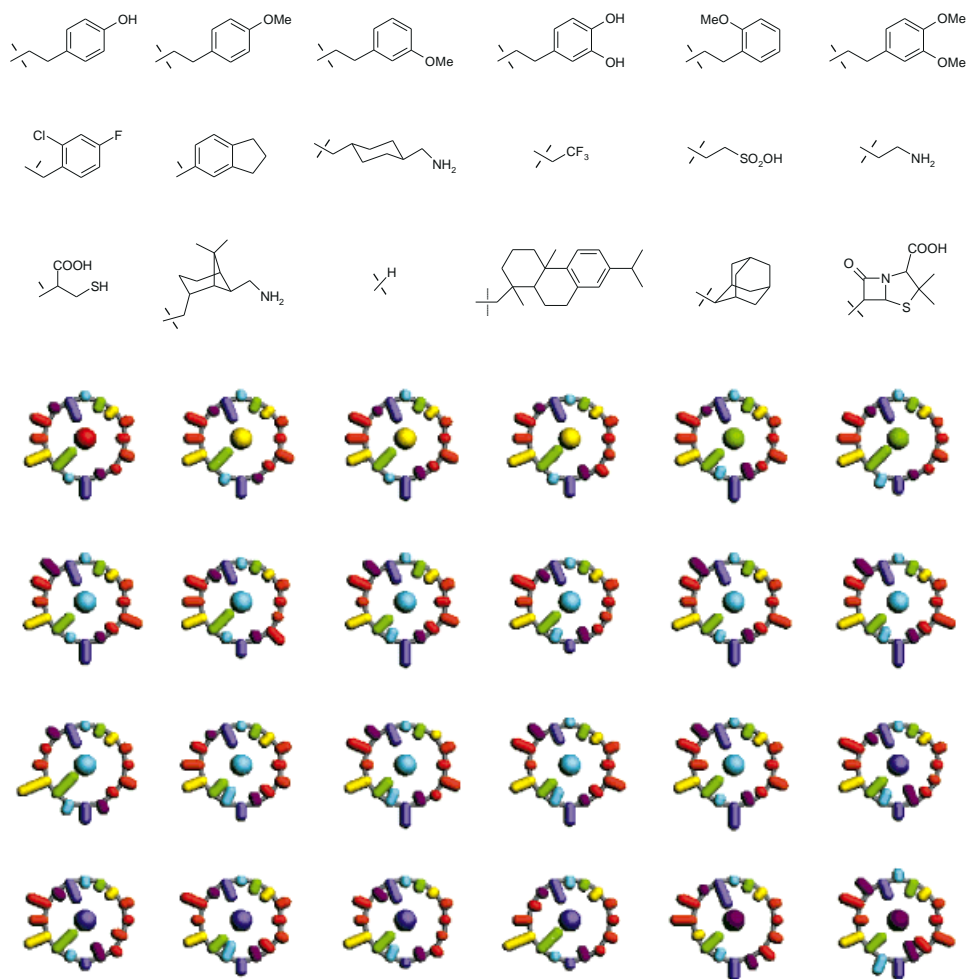


Figure 5 Structures and flower plots of 18 side-chains from a biased *N*-substituted glycine combinatorial library based on the tyramine submonomer (reproduced from *J. Med. Chem.*, 1995, **38**, 1431). The side-chains on the top row include tyramine and its five closest analogs. The 12 side-chains in the lower rows were chosen by D-optimal design from a pool of 721 amines

between the three centers, and each point represents a distinct pharmacophore, labeled by a symbol that indicates the three centers, and color-coded according to whether it contains 1, 2, or 3 identical centers. A typical pharmacophore plot generated using Chemical Design's Chem-X suite is shown in Figure 6.

As we mentioned before (Section 2.1.10), pharmacophore keys may be derived from a single conformation or an ensemble of conformations. Moreover, pharmacophore maps of more than one structure may be combined into a single plot, a technique that is particularly useful for visualizing and comparing the pharmacophoric diversity of combinatorial libraries.

5.5 Self-organizing Maps

Self-organizing maps (SOMs) or Kohonen networks⁶⁴ (see *Neural Networks in Chemistry*) were originally designed in an attempt to model intelligent information processing, i.e., the ability of the brain to form reduced representations of the most relevant facts without loss of information about their inter-relationships. The general idea is to map a set of vectorial samples onto a two-dimensional lattice in a way that preserves the topology of the original space. That is, samples that are similar to each other in the input space should be found 'close' to each other in the output space. SOMs belong to a class of neural networks known as competitive learning or self-organizing networks. All neurons receive identical input, and by means of lateral interactions, they compete in their activities. The main application of the SOM is in visualizing complex data on a two-dimensional array, and in creating abstractions reminiscent of these obtained from clustering methodologies.

A Kohonen network maps a set of n -dimensional data samples onto an ordered collection of neurons which are typically arranged in a rectangular or hexagonal lattice (Figure 7).

Each neuron, i , in the network is associated with a reference vector of weights $m_i = [\mu_{i1}, \mu_{i2}, \dots, \mu_{in}]$. The network is trained in an iterative fashion: each data sample, $x = [\xi_1, \xi_2, \dots, \xi_n]$, is presented to the network in a random order, and its euclidean distance from each neuron is computed. The neuron that is closest to the data sample is the location of the 'response'. The weights, m_i , of the matching neuron (and to a lesser extent those of the neighboring neurons) are then adapted to approach the input sample using a neighborhood function or smoothing kernel defined over the

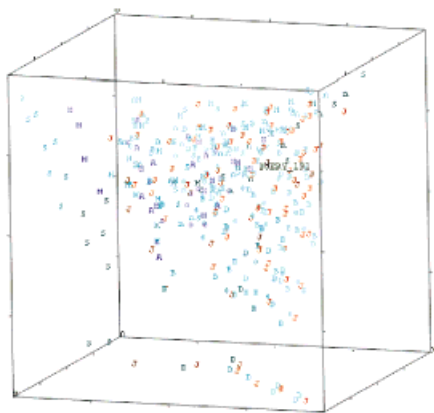


Figure 6 Pharmacophore plot generated by Chemical Design's Chem-X software suite

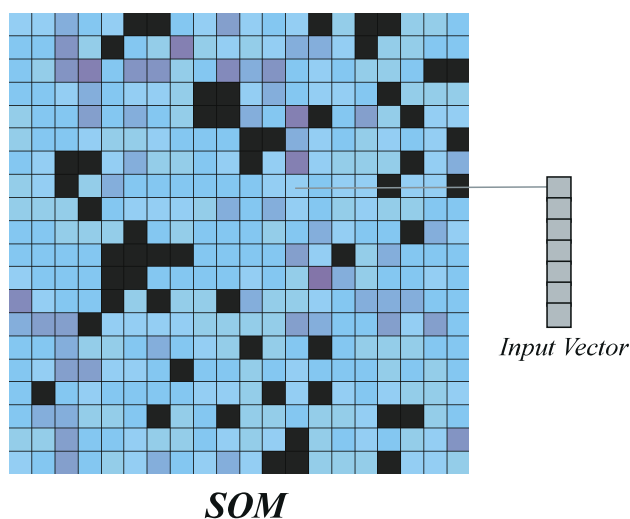


Figure 7 Main organizational principle of a Kohonen map. The input vector is compared to every neuron (cell) in the network, and the one that is closer determines the location of the response

lattice points. A widely used kernel is given in equation (22):

$$h_{ci}(t) = \alpha(t) \exp\left(-\frac{\|\mathbf{r}_c - \mathbf{r}_i\|^2}{2\sigma(t)^2}\right) \quad (22)$$

where c is the matching neuron, \mathbf{r}_c and \mathbf{r}_i are the respective locations of the c -th and i -th neurons on the lattice ($\mathbf{r}_c, \mathbf{r}_i \in \mathcal{R}^2$), $\alpha(t)$ is the learning rate, and $\sigma(t)$ is the width of the kernel. To ensure convergence, it is important that $h_{ci}(t) \rightarrow 0$ as $t \rightarrow \infty$. After repeated application of this training principle, the weight vectors are relaxed and the map becomes globally ordered. Each neuron is sensitized to a different domain of the input space, and acts as a decoder of that domain.

After training is finished, each data point is again presented to the network, and the matching neuron is determined. This process is, in effect, a non-linear projection from an n - to a two-dimensional space. We should point out that SOM is primarily a clustering, visualization, and abstraction method. As in most other techniques of this kind, feature selection is of paramount importance and some preprocessing may be necessary to ensure meaningful results. Although SOMs can be used for classification, there are a number of supervised variants of the basic algorithm such as learning vector quantization (LVQ) and the dynamically expanding context (DEC) that may be more appropriate for this task.⁶⁴

The first application of SOM as a means of analyzing molecular similarity and diversity was reported by Gasteiger and co-workers at the University of Erlangen.²³ Their approach was demonstrated using three combinatorial libraries, originally designed by Rebek as potential inhibitors of the serine protease trypsin. Only the xanthene and cubane libraries were actually synthesized by Rebek. The adamantane library was designed by Gasteiger as a mimic of the cubane library. The libraries were generated by combining the poly-functionalized xanthene, cubane, and adamantane scaffolds shown in Figure 8 with 19 L-amino acids, giving rise to 65 341, 11 191, and 11 191 unique compounds, respectively.

Each virtual compound was represented by a 12-dimensional spatial autocorrelation vector computed according

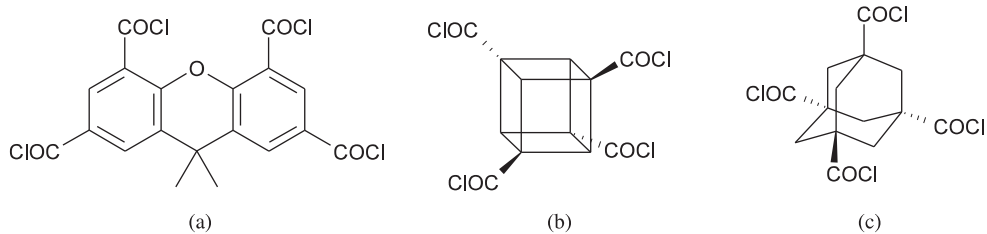


Figure 8 Combinatorial scaffolds used by Sadowski, Wagener, and Gasteiger:²³ (a) xanthene; (b) cubane; (c) adamantane

to equation (23):

$$A(d_l, d_u) = \frac{1}{N} \sum_{i,j} p_i p_j \quad (23)$$

where p_i and p_j are physicochemical property values computed at two randomly chosen points, i and j , on the molecular surface (in this case, the molecular electrostatic potential), d_{ij} is the distance between these points, and N is the total number of distances in the interval $[d_l, d_u]$. These autocorrelation coefficients compress shape and electronic information into a single low-dimensional vector, resulting in a compact, informative molecular descriptor with direct bearing on biological activity.

These descriptors were used to train two Kohonen networks, arranged in a 50×50 rectangular array (Figure 9). The first network was trained using the cubane and xanthene libraries alone, and is shown in Figure 9(a), with each neuron color-coded according to the most frequently occurring core. The cubane library (black) forms a distinct cluster in the middle of the map, surrounded by the xanthene derivatives. Only 3% of the total number of neurons were occupied by members of both libraries, and they were all located on the periphery of the cubane cluster. It is clear that the SOM procedure was able to separate the two classes very effectively, which seems to suggest that the two libraries are structurally diverse and non-redundant.

To prove that SOMs can also be used to assess the similarity of chemical libraries, a second network was trained using an additional virtual library based on the adamantane core functionalized at the four bridgehead positions Figure 8(c). As

shown in Figure 9(b), the SOM again discriminated very effectively between the xanthene and cubane/adamantane derivatives, but was unable to distinguish the cubane from the adamantane library, which is consistent with the conformational constraints imposed by their rigid cores.

As we will see in Section 5.7, this discriminatory ability is not so much due to the SOM procedure itself, but rather the nature of the spatial autocorrelation vector and the severe conformational constraints imposed by the rigid cores used in this study. The usefulness of this approach for comparing non-rigid systems that exhibit some degree of conformational flexibility remains to be seen. Nonetheless, SOMs are not limited to autocorrelation coefficients, and can be applied to any molecular representation of vectorial nature. Given the conceptual simplicity of the output, SOMs can be very effective for visualizing and comparing chemical libraries, particularly when they are coupled with advanced, interactive graphical tools.

5.6 Multi-dimensional Scaling

Multi-dimensional scaling (MDS) emerged from the need to visualize a set of objects described by means of a similarity or dissimilarity matrix. The technique originated in the field of psychology and can be traced back to the work of Torgerson⁶⁵ and Kruskal.⁶⁶ The problem is to construct a configuration of points in a low-dimensional space from information about the distances between these points. In particular, given a set of k data points in the input space $\{x_i, i = 1, 2, \dots, k\}$, a symmetric matrix d_{ij} of the observed dissimilarities between these points, and a set of images of x_i on a d -dimensional

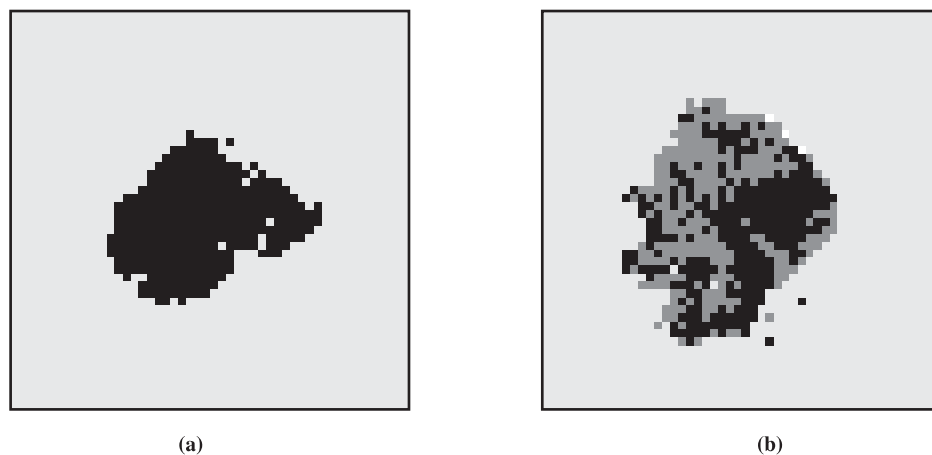


Figure 9 Self-organizing maps of: (a) the xanthene (light gray) and cubane (black) libraries; and (b) the xanthene (light gray), cubane (black), and adamantane (dark gray) libraries

display plane $\{\xi_i, i = 1, 2, \dots, k; \xi_i \in \mathfrak{R}^d\}$, the objective is to place ξ_i onto the plane in such a way that their Euclidean distances $\delta_{ij} = \|x_i - x_j\|$ approximate as closely as possible the corresponding values d_{ij} . A sum-of-squares error function can be used to decide the quality of the embedding. Two commonly used criteria are Kruskal's stress:

$$S = \sqrt{\frac{\sum_{i < j} (\delta_{ij} - d_{ij})^2}{\sum_{i < j} d_{ij}^2}} \quad (24)$$

and Lingoes' alienation coefficient:

$$S = \sqrt{\frac{1 - \sum_{i < j} (\delta_{ij} \cdot d_{ij})^2}{\sum_{i < j} d_{ij}^2}} \quad (25)$$

The actual embedding is carried out in an iterative fashion. The process starts by: (1) generating an initial set of coordinates ξ_i ; (2) computing the distances δ_{ij} ; (3) finding a new set of coordinates ξ_i using a steepest descent algorithm such as Kruskal's linear regression or Guttman's rank-image permutation; and (4) repeating steps 2 and 3 until the change in the stress function falls below some predefined threshold.

MDS can be metric and non-metric. Metric scaling operates on the assumption that the input is either ratio or interval data (quantitative), whereas the non-metric model requires that the data is provided at the ordinal level of measurement, i.e., in the form of ranks (qualitative). Thus, the non-metric model imposes fewer restrictions, but is also less rigorous. Unlike self-organizing maps, MDS does not require vectorial samples and is therefore much more generally applicable.

The first application of MDS in the context of molecular diversity was presented by the group at Chiron⁸ as a means for reducing the enormous dimensionality of binary chemical descriptors. They found that the 2048-bit Daylight fingerprints associated with 721 commercially available primary amines could be reduced to only five continuous variables that reproduced all 260 000 original dissimilarities with a standard deviation of only 10%. Similarly, only seven dimensions were required to reduce the 642 000 pairwise similarities among a set of 1133 carboxylic acids and acid chlorides to the same precision. Although MDS was originally designed as a visualization tool, the cpu requirements of the existing algorithms are prohibitive for visualizing large data sets such as those encountered in combinatorial library designs. This has been successfully achieved using a close relative of MDS known as non-linear mapping.

5.7 Non-linear Maps

Non-linear mapping is a multivariate statistical technique that is closely related to multi-dimensional scaling.⁶⁷ Just like MDS, the objective is to approximate local geometric relationships on a two- or three-dimensional plot.

Non-linear mapping can be metric or non-metric, and is therefore applicable to a wide variety of input data. This is particularly useful when the (dis)similarity measure is not a true metric, i.e., it does not obey the distance postulates and, in particular, the triangle inequality. Although an 'exact'

projection is only possible when the distance matrix is positive definite, meaningful projections can be obtained even when this criterion is not satisfied.

As in MDS, the process starts with a finite set of samples $\{x_i, i = 1, 2, \dots, k\}$, a symmetric dissimilarity matrix d_{ij} , and a set of images of x_i on a display plane $\{\xi_i, i = 1, 2, \dots, k; \xi_i \in \mathfrak{R}^d\}$, and attempts to place ξ_i onto the plane in such a way that their Euclidean distances $\delta_{ij} = \|\xi_i - \xi_j\|$ approximate as closely as possible the corresponding values d_{ij} . The embedding (which can only be made approximately) is carried out in an iterative fashion by minimizing an error function, $E(m)$, which measures the difference between the distance matrices of the original and projected vector sets:

$$E(m) = \frac{\sum_{i < j}^k [d_{ij} - \delta_{ij}(m)]^2 / d_{ij}}{\sum_{i < j}^k d_{ij}} \quad (26)$$

where m is the iteration number. $E(m)$ is minimized using a steepest-descent algorithm. The initial coordinates, ξ_i , are determined at random, and are updated using equation (27):

$$\xi_{pq}(m+1) = \xi_{pq}(m) - \lambda \Delta_{pq}(m) \quad (27)$$

where λ is the learning rate parameter, and

$$\Delta_{pq}(m) = \frac{\partial E(m)}{\partial \xi_{pq}(m)} \bigg/ \left| \frac{\partial^2 E(m)}{\partial \xi_{pq}(m)^2} \right| \quad (28)$$

The first non-linear mapping algorithm was presented by Sammon,⁶⁷ but just like MDS it too is only applicable to relatively small data sets. Our group has developed an alternative self-organizing algorithm which is reminiscent of Kohonen's SOM and neural network back-propagation, and allows the scaling of very large data sets.⁷⁰

Non-linear maps were introduced by Agrafiotis to visualize protein sequence relationships in two dimensions,⁶⁸ and were later employed as a means of visualizing and comparing large compound collections, represented by a set of molecular descriptors.^{44,50} The advantage of Sammon maps compared to Kohonen networks is that they provide much greater detail about the individual compounds and their inter-relationships. A typical output is illustrated in Figure 10. In this example, the data set consisted of 1000 three-dimensional random vectors, 90% of which were distributed normally around three randomly chosen cluster centers, and the remaining 10% were uniformly distributed in the unit cube. The three-dimensional density function is shown in Figure 10(a) and the corresponding non-linear map in Figure 10(b). It is clear that the non-linear projection preserves the topology of the original data set, and reproduces the three clusters in terms of density, spread, and mutual separation.

To provide a direct comparison between self-organized and non-linear maps, we applied our non-linear mapping algorithm on the xanthene, cubane, and adamantane libraries that were used by Gasteiger et al.²³ (Figure 11). The projection was carried out using the same 12-dimensional autocorrelation descriptors and the Euclidean metric as a pairwise measure of dissimilarity. The resulting map is shown in Figure 11.

The map is sufficiently faithful, as manifested by a Sammon and Kruskal stress values of only 10% and 8%, respectively.

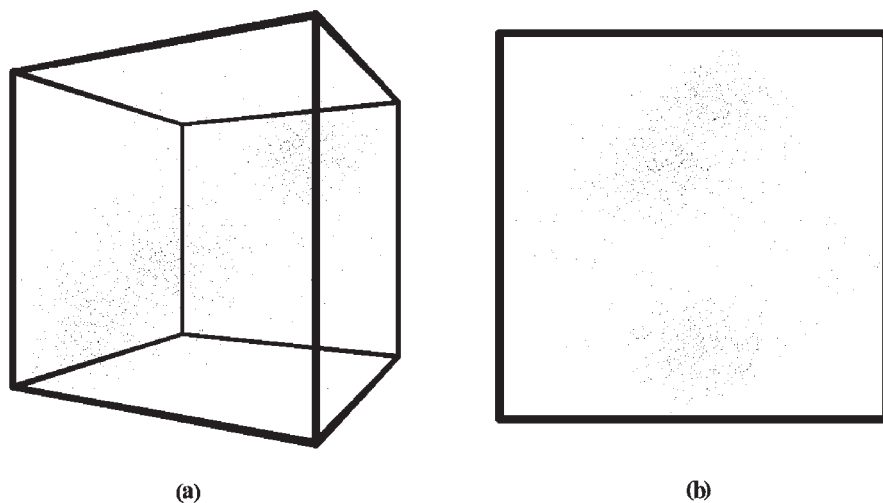


Figure 10 Non-linear map of a 3D data set with three normally distributed clusters; (a) 3D scatter plot; (b) non-linear map

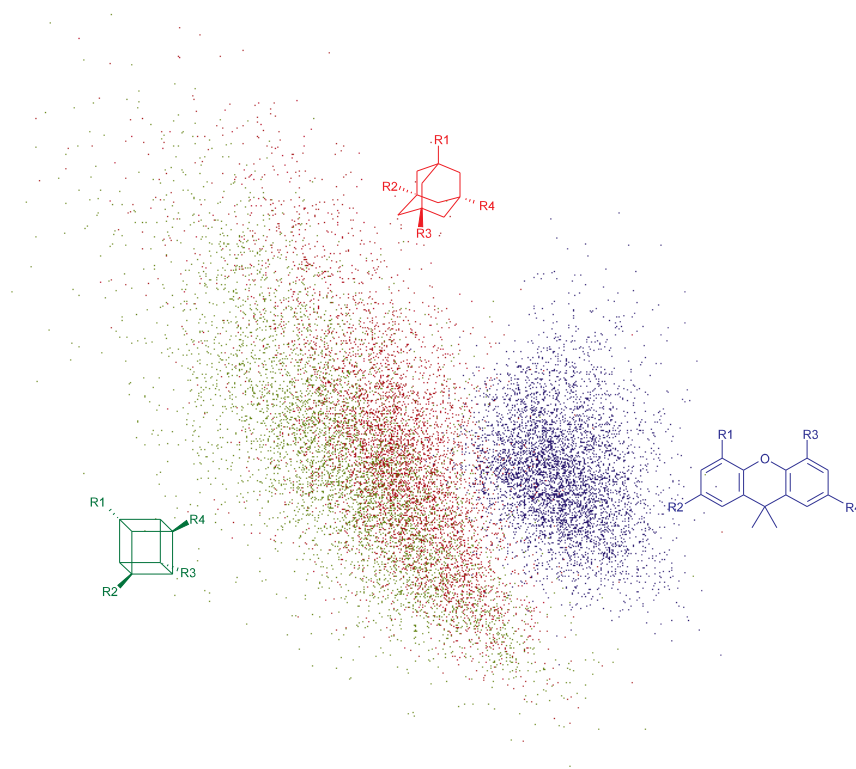


Figure 11 Non-linear map of the xanthene (blue), cubane (green), and adamantane (red) libraries used by Gasteiger et al.

It is evident that the non-linear map is not only capable of reproducing the sharp separation between the planar and tetrahedral systems that was observed in the self-organized maps (Figure 9), but also revealed a more subtle distinction between the cubane and adamantane libraries was not captured by the Kohonen network.

While the first application of this technique involved continuous molecular descriptors, the programs were later extended to include other molecular representations and molecular similarity metrics such as substructure keys, hashed fingerprints, Tanimoto coefficients, etc.⁶⁹

6 CONCLUSIONS

Driven by revolutionary advances in automated synthesis and high-throughput screening, diversity profiling has become an indispensable tool in the hands of the medicinal chemists. It offers the potential of reducing the redundancy and cost of experiments, and can substantially increase the odds of discovering new drugs. Yet, despite its conceptual simplicity, diversity remains an elusive concept that defies a rigorous definition. From a practical point of view, diversity is really a design strategy that attempts to maximize the hit rate of

high-throughput screening experiments. The choice of metrics and descriptors can only be validated to the extent that they meet this deceptively simple goal. Many people believe that diversity is serendipity in disguise, and there are many studies to suggest that this is true. Validation can only come from comparison with appropriate control experiments, but these are hard to design and too expensive to execute. We are clearly at the dawn of an emerging field. Yet, we are guided by decades of experience in the fields of molecular similarity and structure-activity correlation, which provide a fertile ground for developing our theories and approaches. Finally, diversity should never be taken for more than what it is; one should always remember that the optimal series for lead design is neither random nor maximally diverse.

7 RELATED ARTICLES

Chemometrics: Multivariate View on Chemical Problems; Combinatorial Chemistry; Combinatorial Libraries: Structure-Activity Analysis; Computer Graphics and Molecular Modeling; Drug Design; Experimental Design; Genetic Algorithms: Introduction and Applications; Genetic and Evolutionary Algorithms; Graph Theory in Chemistry; High-throughput 'Virtual' Chemistry; Neural Networks in Chemistry; Quantitative Structure-Activity Relationships in Drug Design; Quantitative Structure-Property Relationships (QSPR); Simulated Annealing; Structural Similarity Measures for Database Searching; Structure and Substructure Searching; Structure Databases; Topological Indices; Topological Methods in Chemical Structure and Bonding.

8 REFERENCES

- M. A. Johnson and G. M. Maggiora, 'Concepts and Applications of Molecular Similarity', Wiley, New York, 1990.
- C. Hansch and A. Leo, 'Exploring QSAR. Fundamentals and Applications in Chemistry and Biology', American Chemical Society, Washington, DC, 1995.
- E. J. Martin, D. C. Spellmeyer, R. E. Critchlow, Jr., and J. M. Blaney, in 'Reviews in Computational Chemistry', Vol. 10, eds. K. B. Lipkowitz and D. B. Boyd, VCH, Weinheim, 1997.
- J. M. Blaney and E. J. Martin, *Curr. Biol.*, in press.
- Y. C. Martin, R. D. Brown, and M. G. Bures, in 'Combinatorial Chemistry and Molecular Diversity', eds. J. F. Kerwin and E. M. Gordon, Wiley, New York, 1997.
- P. Willett, 'Similarity and Clustering in Chemical Information Systems', Research Studies Press, Letchworth, 1987.
- G. M. Downs and P. J. Willett, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1094-1102.
- E. J. Martin, J. M. Blaney, M. A. Siani, D. C. Spellmeyer, A. K. Wong, and W. H. Moos, *J. Med. Chem.*, 1995, **38**, 1431-1436.
- R. Lewis, I. M. McLay, and J. S. Mason, *Chem. Design Autom. News*, 1995, **10**(4), 37-38.
- R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 572-584.
- R. D. Brown and Y. C. Martin, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 1-9.
- J. Kubinyi, in 'Methods and Principles in Medicinal Chemistry', Vol. 1, eds. R. Manhold, P. Krosgaard-Larsen, and H. Timmermann, VCH, Weinheim, 1993, pp. 21-36.
- L. B. Kier and L. H. Hall, 'Molecular Connectivity in Structure-Activity Analysis', Wiley, New York, 1986.
- M. Charton and I. Motoc (eds.), 'Steric Effects in Drug Design', Springer-Verlag, Heidelberg, 1983.
- 'Molconn-X', Haney Associates, Mercer Island, WA.
- G. M. Downs and J. M. Barnard, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 59-61.
- R. E. Carhart, D. H. Smith, and R. J. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1985, **25**, 64-73.
- S. K. Kearsley, S. Sallmack, E. M. Fluder, J. D. Andose, R. T. Mosley, and R. P. Sheridan, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 118-127.
- R. P. Sheridan, M. D. Miller, D. J. Underwood, and S. K. Kearsley, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 128-136.
- G. Moreau and C. Turpin, *Analysis*, 1996, **24**, M17-M22.
- H. Bauknecht, A. Zell, H. Bayer, P. Levi, M. Wagener, J. Sadowski, and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1205-1213.
- M. Wagener, J. Sadowski, and J. Gasteiger, *J. Am. Chem. Soc.*, 1995, **117**, 7769-7775.
- J. Sadowski, M. Wagener, and J. Gasteiger, *Angew. Chem., Int. Ed. Engl.*, 1996, **34**, 23-24.
- R. S. Pearlman, *Network Science*, 1996, June, <http://www.awod.com/netsci/issues/>
- F. R. Burden, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 225-227.
- 'Unity Chemical Information Software', Tripos Associates, St. Louis, MO.
- N. W. Murrall and E. K. Davies, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 312-316.
- J. Sadowski, J. Gasteiger, and G. Klebe, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 1000-1008.
- R. P. Sheridan, R. Nilikantan, A. Rusinko, N. Bauman, K. Haraki, and R. Venkataraghavan, *J. Chem. Inf. Comput. Sci.*, 1989, **29**, 255-260.
- S. Pickett, J. S. Mason, and I. M. McLay, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 1214-1223.
- E. K. Davies and C. Briant, *Network Science*, 1995, <http://www.awod.com/netsci/issues/>
- H. Kubinyi, '3D-QSAR in Drug Design: Theory, Methods and Applications', ESCOM Science, Leiden, 1993.
- R. D. Cramer, R. D. Clark, D. E. Patterson, and A. M. Ferguson, *J. Med. Chem.*, 1996, **39**, 3060-3069.
- L. M. Kauvar, D. L. Higgins, H. O. Villar, J. R. Sportsman, A. Engqvist-Goldstein, R. Bukar, K. E. Bauer, H. Dilley, and D. M. Rocke, *Chem. Biol.*, 1995, **2**(2), 107-118.
- D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, and L. E. Weinberger, *J. Med. Chem.*, 1996, **39**, 3049-3059.
- R. E. Bellman, 'Adaptive Control Processes', Princeton University Press, Princeton, 1961.
- D. W. Scott, 'Multivariate Density Estimation: Theory, Practice and Visualization', Wiley, New York, 1992.
- E. J. Wegman, *Ann. Statist.*, 1970, **41**, 457-471.
- S. Teig, in 'Cambridge Healthtech Institute's Conference on Chemoinformatics', May 12-13, Arlington, VA, 1997.
- W. Cooley and P. Lohnes, 'Multivariate Data Analysis', Wiley, New York, 1971.
- S. Gibson, R. McGuire, and D. C. Rees, *J. Med. Chem.*, 1996, **39**, 4065-4072.
- D. J. Cummins, C. W. Andrews, J. A. Bentley, and M. Cory, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 750-763.
- M. Hassan, J. P. Bielawski, J. C. Hempel, and M. Waldman, *Mol. Divers.*, 1996, **2**, 64-74.
- D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.*, 1977, **37**, 841-851.
- M. S. Lajiness, in 'QSAR: Rational Approaches to the Design of Bioactive Compounds', eds. C. Silipo and A. Vittoria, Elsevier, Amsterdam, 1991, pp. 201-204.
- A. Polinsky, R. D. Feinstein, S. Shi, and A. Kuki, in 'Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery', eds. I. M. Chaiken and K. D. Janda, American Chemical Society, Washington, DC, 1996, pp. 219-232.

47. D. J. Chapman, *J. Comput.-Aided Mol. Design*, 1996, **10**, 501-512.
48. R. J. Taylor, *J. Chem. Inf. Comput. Sci.*, 1995, **35**, 59-67.
49. M. Marsili and H. J. Saller, *J. Chem. Inf. Comput. Sci.*, 1993, **33**, 266-269.
50. D. K. Agrafiotis, in '3rd Electronic Computational Chemistry Conference', 1996.
<http://hackberry.chem.niu.edu/ECCC3/paper48>
51. D. K. Agrafiotis, R. F. Bone, F. R. Salemme, and R. M. Soll, United States Pat. 5,463,564, 1995.
52. T. L. Graybill, D. K. Agrafiotis, R. Bone, C. R. Illig, E. P. Jaeger, K. T. Locke, T. Lu, J. M. Salvino, R. M. Soll, J. C. Spurlino, N. Subasinghe, B. E. Tomczuk, and F. R. Salemme, in 'Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery', eds. I. M. Chaiken and K. D. Janda, American Chemical Society, Washington, DC, 1996, pp. 16-27.
53. D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 576-580.
54. D. K. Agrafiotis, *J. Chem. Inf. Comput. Sci.*, submitted.
55. N. E. Shemetulskis, D. Weininger, C. J. Blankley, J. J. Yang, and C. J. Humblet, *J. Chem. Inf. Comput. Sci.*, 1996, **36**, 862-871.
56. S. M. Boyd, M. Beverly, L. Norskov, and R. E. Hubbard, *J. Comput.-Aided Mol. Design*, 1995, **9**, 417-424.
57. P. A. Bartlett, 'Abstracts of Papers of the American Chemical Society', 1996, p. 211.
58. S. K. Lin, *Molecules*, 1996, **1**, 57-67.
59. D. B. Turner, S. M. Tyrrell, and P. Willett, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 18-22.
60. H. J. Chernoff, *Am. Statist. Assoc.*, 1973, **68**, 361-368.
61. D. F. Andrews, *Biometrics*, 1972, **28**, 125-136.
62. S. E. Fienberg, *Am. Statist.*, 1979, **33**, 165-178.
63. A. Inselberg, *The Visual Computer*, 1985, **1**, 69-91.
64. T. Kohonen, 'Self-Organizing Maps', Springer-Verlag, Heidelberg, 1996.
65. W. S. Torgerson, *Psychometrika*, 1952, **17**, 401-419.
66. J. B. Kruskal, *Psychometrika*, 1964, **29**, 115-129.
67. J. W. Sammon, *IEEE Trans. Comput.*, 1969, **C-18**, 401-409.
68. D. K. Agrafiotis, *Protein Sci.*, 1997, **6**, 287-293.
69. D. K. Agrafiotis and V. S. Lobanov, unpublished results.
70. Patents pending.