

# Visualization of the Chemical Space in Drug Discovery

Jose L. Medina-Franco<sup>\*1</sup>, Karina Martínez-Mayorga<sup>1</sup>, Marc A. Giulianotti<sup>1</sup>, Richard A. Houghten<sup>1,2</sup> and Clemencia Pinilla<sup>2</sup>

<sup>1</sup>Torrey Pines Institute for Molecular Studies, 5775 Old Dixie Highway, Fort Pierce, FL 34946, USA

<sup>2</sup>Torrey Pines Institute for Molecular Studies, 3550 General Atomics Court, San Diego, CA 92121, USA

**Abstract:** Chemical space has become a key concept in drug discovery. The continued growth in the number of molecules available raises the question regarding how many compounds may exist and which ones have the potential to become drugs. Analysis and visualization of the chemical space covered by public, commercial, in-house and virtual compound collections have found multiple applications in diversity analysis, *in silico* property profiling, data mining, virtual screening, library design, prioritization in screening campaigns, and acquisition of compound collections, among others. This review covers several techniques, computational programs and approaches that have been developed to visualize, navigate and study the chemical space of molecular databases. Techniques developed in our group are presented including a quantitative assessment of the multi-fusion similarity maps. Additionally an application of 3D-similarity, based on the overlay of chemical structures, to represent the chemical space is introduced. Several comparisons of the chemical space covered by compound collections from different sources such as combinatorial libraries, drugs and natural products, or directed to specific therapeutic areas are also discussed.

**Keywords:** Chemoinformatics, combinatorial libraries, data-driven analysis, data mining, molecular diversity, multi-fusion similarity maps, structure-activity relationships, virtual screening.

## INTRODUCTION

Drug discovery programs nowadays frequently involve a vast amount of data. For example, combinatorial chemistry and high-throughput screening generate large outputs that are stored in public, in-house or commercial databases [1, 2]. The continued growth in the number of molecules stored in these databases raises the question of how these molecules compare to each other and to the theoretical number of chemical structures (see below) in terms of number and diversity. A related question frequently asked when designing and screening new libraries [3] is concerned with the new libraries potential therapeutic interest with respect to currently known compounds. In order to conceptualize the total number of molecules, either real or virtual, the concept of *chemical space* or *chemical universe* is frequently used as an analogy to the cosmic universe. There are other terms that can be found in the literature that refer to the same space, for example *multi-dimensional descriptor space*, *molecular-diversity-* and *property space*. Other terms that will be discussed later in this work are *biologically active space*, *binding site-based chemical space*, *druglike-*, *medicinal chemistry-*, *pharmacological space* and *receptor relevant subspace*.

Computational methodologies aimed at analyzing large amounts of data relevant to drug discovery have been reviewed elsewhere [4-8]. In this paper we review techniques aimed at obtaining a visual representation of chemical space as well as practical applications and conclusions derived from the visual comparisons of compound collections.

This review is divided into a few sections. The importance of visualization in drug discovery and common computational techniques for representing chemical space are discussed first. Then several applications of chemical space visualization using different approaches are presented. Finally a brief example describing the combination of chemical space visualization with structural analysis is presented before the conclusion.

## WHAT IS CHEMICAL SPACE?

Despite the widespread use of the term *chemical space* there are few definitions that have been proposed for it. In a direct comparison with the cosmic space, Lipinski and Hopkins state that “chemical space can be viewed as being analogous to the cosmological universe in its vastness, with chemical compounds populating space instead of stars” [9]. Dobson in his insight defines chemical space as “the total descriptor space that encompasses all the small carbon-based molecules that could in principle be created” [10]. At a previous Horizon Symposia, entitled “Charting chemical space: finding news tools to explore biology”, chemical space was described as “the set of all possible molecular structures” [11, 12]. As it will be discussed in this review, the representation of the chemical space of a compound collection may vary with the particular set of descriptors and parameters used to define the space where the molecules will be located.

## HOW BIG IS THE CHEMICAL SPACE?

It is widely accepted that the chemical space is huge and that there is actually only a small fraction of molecules that are known. A yet smaller fraction of compounds seem to be relevant for medicinal chemistry purposes [11]. Different estimates have been proposed for the size of chemical space. For example Petit-Zeman describes the number of molecules to be between  $10^{18}$  and  $10^{200}$  [12]. Geysen *et al.* [13] mention

\*Address correspondence to this author at the Torrey Pines Institute for Molecular Studies, 5775 Old Dixie Highway, Fort Pierce, FL 34946, USA; Tel: +1-772-462-0891; Fax: +1-772-462-0886; E-mail: jmedina@tpims.org

that the number of small molecules could be between  $10^{14}$  and  $10^{30}$ . Bohacek *et al.* estimated the number of compounds with a maximum number of 30 C, N, O and S atoms to be  $10^{60}$  [14]. Of course, this estimate will increase when considering larger and more complex structures. To illustrate this point, Dobson calculated the number of putative proteins containing 300 residues, average size of a natural protein, considering 20 different types of amino acids as more than  $10^{390}$  ( $20^{300}$ ). Ertl suggest that the organic chemistry space contains between  $10^{20}$  and  $10^{24}$  molecules synthetically feasible using currently known synthetic methods [15]. Starting from three-dimensional structures of ligands obtained from the Protein Data Bank (PDB) [16], Ogata *et al.* [17] generated between  $10^8$  and  $10^{19}$  compounds by considering all possible combinations of atomic species. Atomic species were composed of C, N, O, S, Cl and different numbers of H atoms [17]. Fink *et al.* assembled a database containing all molecules up to 11 atoms of C, N, O, and F, considering simple valence, chemical stability and synthetic feasibility. This database contains 26.4 million compounds and 110.9 million stereoisomers [18].

### VISUALIZATION IN DRUG DISCOVERY

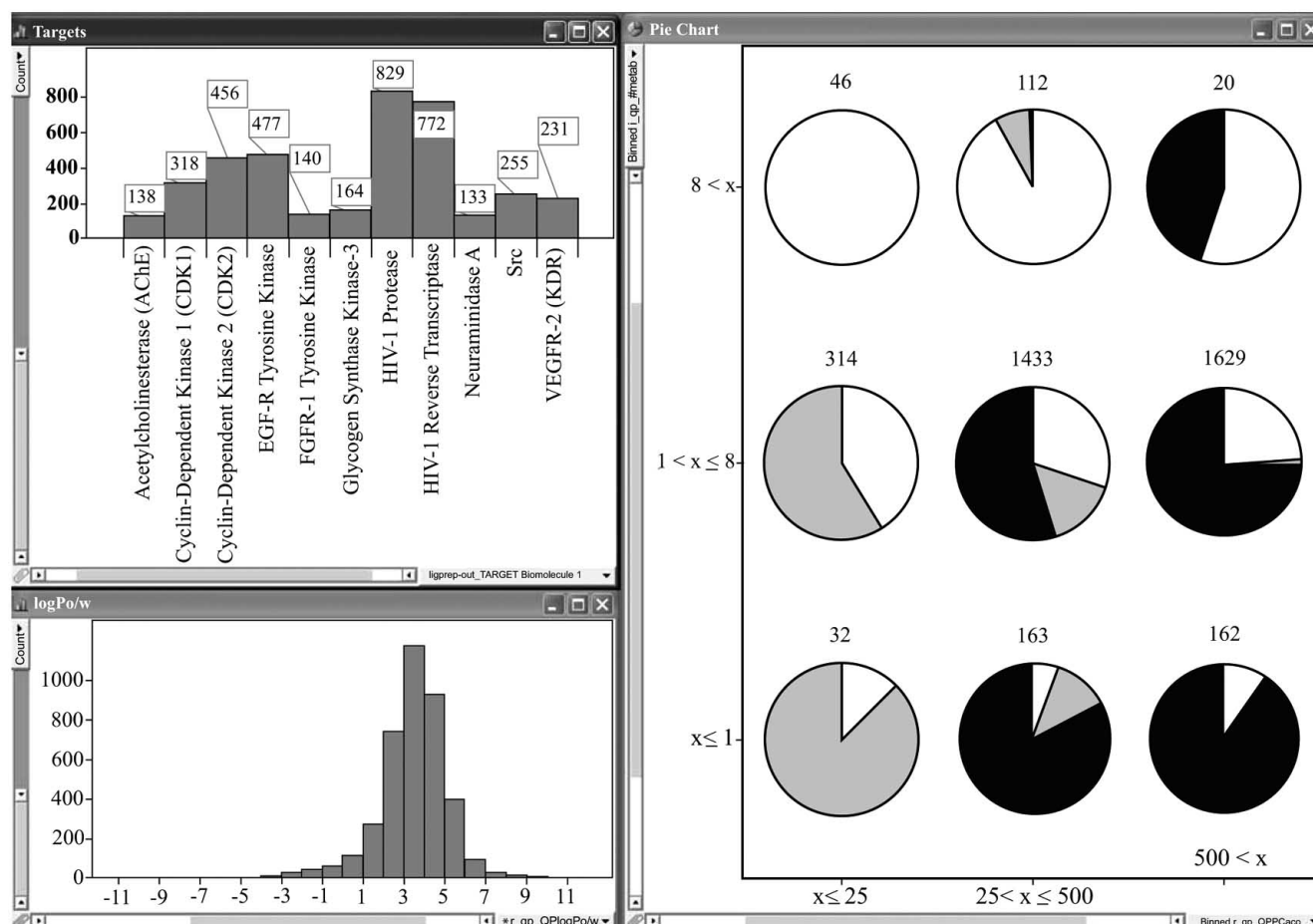
In the current drug discovery process either at the level of hit identification or lead optimization stage, it is common to deal with multivariate data sets with high complexity [19]. This is the result of the increasing advances in the fields of combinatorial chemistry, high-throughput screening and the multiple read outs that can be obtained from *in silico* and phenotypic screening, functional genomics, adsorption, distribution, metabolism, excretion and toxicity (ADMET) profiling [19, 5]. Examples of publicly available databases that may contain thousands or even millions of compounds for analysis have been reviewed recently [1, 2]. To better under-

stand the underlying information contained in multivariate data sets, dimensionality reduction techniques as well as the development of visualization methods have been the focus of intense research [19-21]. Data visualization has been widely recognized as a useful tool to obtain information from large quantities of data [20, 22, 23]. In this context, the term *visualization* can be understood as any methodology that projects data in lower-dimensional space, usually two or three dimensions, while maintaining a large percentage of the information from the higher-dimensional space [20, 24]. In this scenario, applications of data visualization in drug discovery include, but are not limited to, the prioritization of molecules in a compound library for synthesis or biological evaluation [20]; diversity analysis [24-26]; analysis of high content screening data [27]; structure-activity relationships [28]; virtual screening [29, 26] and ADME profiling [30]. A number of computer programs and tools are available with powerful graphics for data visualization. Several commercial and publicly available resources are summarized in Table 1. Some of these programs include algorithms to perform clustering and conduct a dynamic analysis of the data.

An example of visualizing the ADME profiling of the Binding Database [31] is illustrated in Fig. (1). The figure shows the results for 3,913 active compounds (i.e.,  $IC_{50} \leq 1000$  nM, as annotated in the Binding Database) directed to 11 targets. Molecules were prepared with LigPrep [32] and ADME-related properties or descriptors were computed with QikProp [33]. The histogram at the upper left of Fig. (1) shows the distribution of the 3,913 compounds over the targets. The histogram at the lower left shows the distribution of calculated LogP. The pie chart displays three computed properties namely apparent Caco-2 cell permeability, number of likely metabolic reactions and human oral absorption. The

**Table 1. Examples of Commercial and Publicly Available Resources for Visualization and Analysis of Chemical Space**

Program / Tool	Description	Web Site
Spotfire	Suite of applications for interactive visualization and data analysis	<a href="http://spotfire.tibco.com">http://spotfire.tibco.com</a>
Partek	Software for visualization and data analysis with statistical tools	<a href="http://www.partek.com">http://www.partek.com</a>
Miner3D	Integrated data-driven 3D visualization and data analysis program	<a href="http://www.miner3d.com">http://www.miner3d.com</a>
SciTegic Pipeline Pilot	Platform for retrieval, filtering, and data analysis	<a href="http://accelrys.com">http://accelrys.com</a>
DiverseSolutions	Suite of tools to visualize, compare and select libraries	<a href="http://www.tripos.com">http://www.tripos.com</a>
PartiView	Open source interactive viewer for 4-dimensional datasets	<a href="http://viridir.ncsa.uiuc.edu/partiview">http://viridir.ncsa.uiuc.edu/partiview</a>
InfVis	Visual data mining tool	<a href="http://www2.chemie.uni-erlangen.de/research/information_visualization">http://www2.chemie.uni-erlangen.de/research/information_visualization</a>
VlaaiVis	Tool for visualization of structure-activity relationships	<a href="http://www.vlaaivis.com">http://www.vlaaivis.com</a>
Ingenuity Pathways Analysis	Program for the searching, visualization, and analysis of targets, biomarkers, and biological functions	<a href="http://www.ingenuity.com/products/pathways_analysis.html">http://www.ingenuity.com/products/pathways_analysis.html</a>
ChemSpaceShuttle	Graphical interface based on linear and non-linear projection techniques and clustering algorithms	<a href="http://gecco.org.chemie.uni-frankfurt.de">http://gecco.org.chemie.uni-frankfurt.de</a>
Interactive SOM to mine the NCI Antiviral Screen	SOM of approximately 42,000 compounds in the NCI Antiviral Screen	<a href="http://cactus.nci.nih.gov/services/som_qsar">http://cactus.nci.nih.gov/services/som_qsar</a>



**Fig. (1).** Visualization of computed ADME profile of 3,913 molecules from the Binding Database [31] with the program Spotfire DecisionSite 9.1.1. [34]. Histograms depict the distribution of the molecules over 11 targets (top) and distribution of LogP (bottom). The pie charts show the computed apparent Caco-2 cell permeability (nm/sec) as binned values on the X-axis, and number of likely metabolic reactions as binned values on the Y-axis. Charts are further distinguished by the predicted qualitative human oral absorption as low (white), medium (gray) and high (black). Each pie chart has a number at the top indicating the number of compounds.

number at the top of each pie chart indicates the number of molecules in that chart. The X-axis represents binned values of the predicted apparent Caco-2 cell permeability in nm/sec. Values lower than 25 nm/sec mean poor permeability whereas values greater than 500 nm/sec mean great permeability [33]. The Y-axis indicates the number of likely metabolic reactions, also binned in three categories. In this case, it is not desirable to have more than 8 metabolic reactions [33]. The pie charts are shaded by the predicted qualitative human oral absorption as low (white), medium (gray) and high (black). The visualization depicted in Fig. (1) is useful to rapidly capture the profile of ADMET-related properties of the data set. The dynamic nature of some visualization programs (Table 1) helps to quickly explore specific relationships in the data.

As discussed above, the analysis of the chemical space of a compound collection is typically a multi-dimensional descriptor problem where visualization techniques are required to project all compounds in a lower dimensional space, amenable to analysis by the human eye. Several molecular representations (e.g., descriptors, properties, etc.) and dimension reduction techniques have been used to visualize

the chemical space and extract useful information. These techniques along with applications of the techniques are discussed in the following section.

## COMPUTATIONAL APPROACHES TO VISUALIZE THE CHEMICAL SPACE

Visualization of the chemical space of a compound collection will largely depend on two main factors: the molecular representation of the molecules to define the multi-dimensional descriptor space (*vide supra*), and the visualization technique used to reduce the multi-dimensional space into a two- or three-dimensional graph. Noteworthy, the chemical space of a compound collection will not be unique as it will depend on the particular representation used to define the multi-dimensional space [35]. In other words, changes in chemical representation are likely to change neighborhood relationships in chemical space [36, 37] producing a “lack of invariance of chemical space” [36].

Several visualization techniques used to explore structure-activity relationships, conduct structural analysis and other applications in drug discovery have been summarized by Agrafiotis [28] and Maniyar [20]. A number of multi-

mensional data mining tools are available such as hierarchical clustering, decision trees, multidimensional scaling, genetic algorithms, neural networks and support vector machines. Two common visualization methods used so far to represent chemical space are principal component analysis (PCA) and self-organizing maps (SOMs) also known as Kohonen networks. As it will be shown later in multiple examples, many studies have been conducted using PCA or SOMs or in some cases both techniques. Other visualization techniques that will be discussed are multi-fusion similarity (MFS) maps, radar plots, Sammon mapping, activity-seeded structure-based clustering, singular value decomposition, minimal spanning tree, k-means clustering, generative topographic mapping (GTM) and hierarchical GTM.

Briefly, PCA is a linear projection technique in which data vectors are projected into lower dimensions called principal components [38]. The principal components are linear combinations of the original variables. Specifically, the principal components are the eigenvectors of the variance-covariance matrix of the original matrix (i.e., initial multidimensional space). The first principal component corresponds to the largest eigenvalue and explains the largest amount of the variance, that is, the largest amount of information from the initial variables. The second principal component corresponds to the second largest eigenvalue and so on. The eigenvectors (i.e., principal components) are orthogonal to each other. The underlying application of this approach in data visualization is that much of the variation in the data set of the original multi-dimension space can be frequently explained by two or three principal components. The low number of principal components can be represented in a two- or three-dimensional coordinate system. The relative distance that exists between compounds in the PCA derived chemical space becomes a measure of their similarity with respect to the particular molecular representation used [39].

In contrast to PCA, SOM is a nonlinear multidimensional mapping tool that belongs to the artificial neural network methods [40-42]. In SOMs a set of objects is typically mapped into a rectangular array of nodes. Similar objects are mapped into the same or proximal nodes. In contrast, dissimilar objects map into distant nodes [28]. Each neuron is assigned a number of weights that correspond to the number of input variables (i.e, descriptors). In the learning stage of a Kohonen network, the values of the weights in the nodes are first assigned as random numbers. Then, a molecule of the data set is projected into the neuron that has the closest weight values to the input variables of the molecule. This neuron is named the winning neuron. In the following steps the weight of the winning neuron and neighboring neurons are updated. After the adjustments of weights, a second molecule from the data set is taken and a single neuron is selected as a winner, the weights are then adjusted and the process is repeated until all molecules in the data set are assigned to a specific neuron [43].

In the next two sections applications of the techniques mentioned above for visualizing the chemical space of compound collections are described. Table 2 summarizes representative examples.

## EXPLORING DIVERSITY IN CHEMICAL SPACE

### Comparing Data Sets from Multiple Sources

Library selection and design, screening campaigns and virtual screening are examples of tasks in drug discovery that often require the comparison of data sets from multiple sources such as combinatorial chemistry, known drugs, natural products, target oriented compound collections, and diverse data sets. Representative examples of the chemical space covered by compound databases from diverse sources are reviewed next.

#### *Drugs, Natural Products, Combinatorial Libraries and Other Sets*

Using PCA, a two-dimensional representation of the chemical space was generated for three compound collections containing 3,287 natural products, 10,968 drug molecules and a random selection of 13,506 molecules obtained from combinatorial libraries [44]. Molecules were represented using ten descriptors including the number of C-N, C-O, C-S, C-halogen bonds, number of chiral centers and rotatable bonds, ratio of aromatic atoms to ring atoms, ring fusion degree, number of hydrogen-bond donors and number of hydrogen-bond acceptors. Visualization of the chemical space along with analysis of the descriptors distribution provided valuable insights for characterizing the diversity of the three different data sets. It was concluded that natural products and compounds obtained from combinatorial libraries have very different properties. It was found that combinatorial compounds are significantly less diverse than natural products and drug molecules, and that compounds from combinatorial libraries occupy a region of the chemical space where natural products show low diversity [44]. In a similar study, the chemical space of natural products and drugs was visualized using SOMs [55].

In a separate study, Shelat *et al.* [3] used PCA and radar plots to compare the chemical space of six different types of screening libraries including 12,361 natural products, 4,749 bioactive molecules, 33,178 molecular fragments, 86,246 structures satisfying the Lipinski's Rule of 5 [56], 15,060 molecules from diversity-oriented synthesis (DOS) and 7,958 drugs. Chemical structures were represented by molecular properties such as molecular weight (MW), octanol/water partition coefficient (logP), number of hydrogen-bond donors and acceptors and polar surface area, among others. It was concluded that molecules from diversity-oriented synthesis cover a different region of the chemical space [3].

#### *Drugs, Target-Focused and Diverse Sets*

The chemical space of four compound collections including 1,055 compounds from DrugBank, 1,990 molecules from NCI Diversity, a collection of 77 FDA-approved psychotic drugs and a set of 196 active molecules in a kappa opioid receptor binding assay obtained from World of Molecular Bioactivity (WOMBAT) was compared using PCA [26]. Molecules were represented with MDL keys fingerprints. The similarity matrix, constructed using the Tanimoto coefficient, was then subjected to PCA and the first three principal components were used to generate a

**Table 2. Representative Approaches to Visualize the Chemical Space of Different Data Sets and Applications**

Data Sets	Molecule Representation	Visualization Technique	Application	Ref.
Natural products, drugs and compounds from combinatorial libraries	Topological properties	PCA	Diversity analysis	[44]
Natural products, drugs, bioactive molecules, Lipinski's rule of 5 compliant, compounds from DOS and molecular fragments	Topological and physico-chemical properties	Radar plots and PCA	Diversity analysis	[3]
Drugs, bioactive molecules and diverse set	Binary fingerprints	Multi-fusion similarity maps	Diversity analysis and development of MFS maps	[26]
Combinatorial libraries	Spatial autocorrelation descriptors	SOMs	Diversity analysis	[45, 46]
Combinatorial libraries	Binary fingerprints	Multi-fusion similarity maps	Diversity analysis and library selection	This work
Combinatorial libraries	Shape-based 3D-similarity	PCA	Diversity analysis	This work
Natural products	Topological and physico-chemical properties	PCA	Development of ChemGPS-NP. Biologically relevant space	[39]
MDDR	Two-dimensional topological fingerprints and 3D structural fingerprint	Branched tree obtained with a minimal spanning tree	Development of drug-target networks	[47,48]
AChE inhibitors, MDDR and combinatorial libraries	BCUT descriptors selected with activity-seeded structure-based clustering	Two- and three-dimensional plots	Development of receptor-relevant subspace	[49]
GPCR and kinase ligands	Physicochemical properties	Generative topographic mapping (GTM) and hierarchical GTM	Explore large data sets	[20]
Compounds with anti-cancer activity	Two-dimensional molecular descriptors	PCA	Diversity analysis	[50]
Estrogens, monoamino oxidase inhibitors, pesticides and other sets	Topological and binary descriptors	Singular value decomposition and minimization procedure	Diversity analysis and design	[51]
Compounds with anti-AIDS activity	Topological pharmacophore descriptors	Diverse algorithms	Development of ChemSpaceShuttle	[52]
Orally active compounds	VolSurf descriptors	TS-SOM, Sammon mapping, U matrix and k-means clustering	Diversity analysis	[53]
Toxic and nontoxic	Proprietary fingerprint	SOMs	Diversity analysis and classification studies	[54]

three-dimensional representation of the chemical space. The distribution of molecules in the chemical space was in agreement with the individual similarities of the compound collections. The three-dimensional representation for the above described data sets and for a small set of HIV reverse transcriptase inhibitors obtained from the Binding Database were used as a reference to develop MFS maps [26]. MFS maps have been demonstrated to be a very useful technique to visualize the chemical space in two dimensions and to extract quantitative information about the relationships in data sets (*vide infra*). Following a similar methodology the chemical space of a set of 48 compounds obtained from a combinatorial library has been visualized [57].

### Diversity and Size of Chemical Space with Virtual Collections

In order to explore the chemical space of small organic molecules and evaluate the potential of finding new drugs,

large collections with virtual compounds have been assembled. The chemical space of the virtual libraries has been compared to space of known molecules.

Fink *et al.* constructed a virtual library containing 13.9 million compounds with low MW (<160 Daltons) [58]. To obtain a representation of the chemical space of this data set, molecules were described by means of topological and physicochemical properties relevant to drugs such as MW, logP, number of hydrogen-bond donors and acceptors, fraction of rotatable bonds, and topological polar surface area (TPSA). The multi-dimensional space was reduced with PCA and visualized using the first two principal components. The chemical space of the virtual library was compared with the space covered by a reference data set with 36,227 known compounds. The reference set was assembled from commercial and publicly available compound libraries and the size of the molecules were limited to those containing no more than 11 main atoms. The comparison revealed that the virtual

library covers the chemical space more densely than the reference set, and that it populates regions of high-polarity not occupied by the reference set [58]. The study supports the notion that there are several small organic molecules still unknown but synthetically accessible, covering a wide range of properties that may lead to bioactive compounds. The chemical space was also used to visualize results of a virtual screening experiment for G-protein coupled receptors (GPCRs), kinases and ion channels [58].

More recently, Fink *et al.* assembled a virtual library of synthetically accessible molecules containing no more than 11 atoms of C, N, O and F. The library contains 26.4 million molecules (*vide supra*) [18]. The chemical space of the virtual collection was visualized using PCA and SOMs. For PCA, molecules were described with six topological and physicochemical properties namely MW, logP, number of hydrogen-bond donors and acceptors, number of rotatable bonds, and TPSA. The chemical space was visualized using the first two principal components. Similar to the comparison described above for the 13.9 million compound set, the virtual library was compared with a reference collection of known compounds concluding that the virtual library contains compounds in high-polarity regions not occupied by the known molecules. In order to build the SOMs autocorrelation descriptors were used. Several different properties were mapped into the resulting SOMs including structural types, presence or absence of chirality and lead-likeness [18]. Additionally, results of a virtual screening experiment for GPCRs, kinases and ion channel modulators were mapped into the SOMs. More recently this group has developed a methodology to travel through chemical space [59].

In a separate work, Ogata *et al.* [17] generated between  $10^8$  and  $10^{19}$  compounds starting from three-dimensional structures of ligands (*vide supra*). In this study, the chemical space of molecules of different set sizes, i.e., molecules selected at random, molecules with non-leadlikeness filters and molecules from commercially available chemical libraries were compared using PCA. MDL keys and molecular properties including MW, logP, solvent accessibility surface areas, molecular refractivity, and others, were used to represent the molecules. This comparison shows that the chemical space of known compounds is smaller than the space of the virtually generated structures [17].

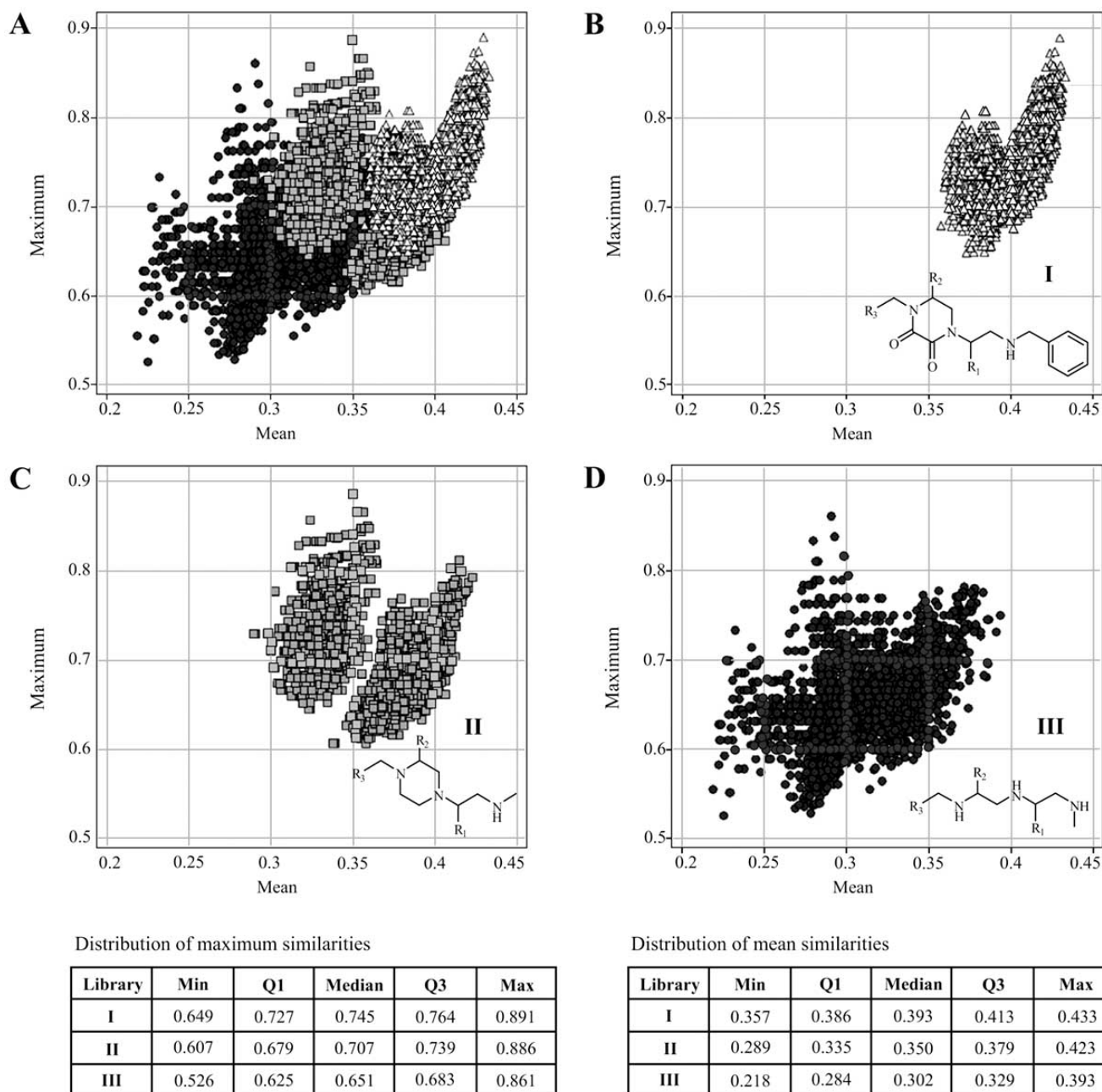
### Combinatorial Libraries

The chemical space of combinatorial libraries has been visualized by means of SOMs and spatial autocorrelation descriptors [45, 46]. One example is the chemical space of three combinatorial libraries with the scaffolds xantane, cubane and adamantane. The three libraries contained more than 87,000 compounds together. The SOMs showed a good separation of the xantane and cubane libraries in chemical space. The cubane and adamantane libraries, however, could not be distinguished in the SOMs, covering a similar region of the chemical space [45, 46].

In a separate study, the chemical space of 7,200 small molecules from a combinatorial library with four points of diversity was visualized using PCA [60]. Molecules were described with MW, computed LogP, atom and bond type counts, electronegativity, connectivity, shape, and electrotopological descriptors. In order to analyze contributions to

diversity, the four substitution positions were mapped into the chemical space and structure-activity relationships were performed for deacetylase inhibitory activity [60]. A similar methodology using PCA was employed to represent the chemical space of the 125 compounds in a benzodiazepine combinatorial library with three points of diversity [61]. Molecules were represented using a number of descriptors including topological indices. Multiple combinatorial libraries have also been visualized in BCUT-derived diversity spaces using a cell-based methodology [25].

MFS maps (*vide supra*) is a recent method proposed for the visual characterization and comparison of compound databases. MFS maps are based on data-fusion similarity measures and characterize the relationship of test molecules (e.g., combinatorial libraries), to a set of reference molecules. The approach, explained extensively elsewhere [26], has recently been used to explore structure-activity relationships of compounds obtained from mixture-based combinatorial libraries [62]. Here we show an implementation of this method in order to compare combinatorial libraries. In Fig. (2) the comparison of three combinatorial libraries (test molecules), with a set of known 1,055 drugs (reference molecules) obtained from DrugBank [63] is depicted. Each library is made up of 31,320 compounds and has three points of diversity. The libraries, denoted here as **I-III**, have identical side chain functionalities at each position ( $R_1 = 29$  reagents,  $R_2 = 27$  reagents and  $R_3 = 40$  reagents). Thus, the libraries differ only in the chemical nature of the central scaffold that is also depicted in Fig. (2): **I** (*N*-benzyl-1,4,5-trisubstituted-2,3-diketopiperazine), **II** (*N*-benzyl-1,4,5-trisubstituted-2,3-piperazine) and **III** (*N*-methyl triamine) [57, 64]. Fig. (2A) portrays the MFS map comparing libraries **I-III** with molecules from DrugBank. The X-axis indicates the *mean* similarity of each molecule in a given combinatorial library with all molecules in DrugBank. The Y-axis denotes the similarity value of each molecule in a given combinatorial library with the most similar molecule in DrugBank (i.e., the *maximum* similarity). Note that the reference set is not shown explicitly in the map [26]. To compute the similarities, molecules were represented with MDL keys [65] as implemented in the Molecular Operating Environment (MOE) program [66] and the similarity was computed with the Tanimoto coefficient [67]. For clarity, Fig. (2B-D) display the libraries in separate maps. From the MFS maps it is readily observed that, out of the three combinatorial libraries, **I** has, in general, the largest mean and maximum similarity values with respect to the known drugs. As opposed to **I**, library **III** has the lowest mean and maximum values. This means that, out of the three libraries, library **III** is the most dissimilar to drugs collected in DrugBank. Furthermore, the relatively low similarity values of molecules in library **III** with respect to known drugs, particularly the maximum similarities, indicates that library **III** is a good candidate to explore other regions in chemical space not covered by known drugs. In order to further quantify the MFS maps, the distribution of the mean and maximum similarity values is summarized at the bottom of the figure. The corresponding maximum, third quartile, median, first quartile and minimum values of the two distributions further support the above conclusion that library **III** is the most dissimilar to the reference and is a good candidate to expand the chemical space covered by known drugs represented in DrugBank. Interest-



**Fig. (2).** Multi-fusion similarity map comparing combinatorial libraries **I-III** with 31,320 compounds each, and known drugs from DrugBank. The reference set (DrugBank) is omitted in the map (see text for details). Panel **A** displays the three libraries and panels **B-D** depicts the libraries separately. The bottom tables indicate the minimum (Min), first quartile (Q1), median, third quartile (Q3) and maximum-similarity distribution of each library.

ingly, library **II** is arranged into two well defined clusters as an effect of the presence of hydroxyl groups at the  $R_1$ ,  $R_2$  and  $R_3$  positions. Compounds in the lower right cluster of library **II** have a hydroxyl group at either,  $R_1$ ,  $R_2$  or  $R_3$  substituent. Noteworthy, the MFS map approach can be applied to virtually any drug discovery program by choosing the appropriate reference and test molecules [26]. For example, for a virtual screening application the reference collection could be a set of active compounds in a particular assay and the test set could be an in-house or commercially available database. An alternative method for conducting a quanti-tative

comparison of combinatorial libraries is the Diversity Space methodology [68, 69].

### VISUALIZING FOCUSED CHEMICAL SPACES

Several efforts have been made to classify small molecules according to their target types. Following the assumption that similar molecules have similar activity [70, 71], the *biologically active space* or *biochem space* [24] would be formed, at least in principle, by separated clusters of compounds each one associated with a different receptor [9]. However, it is not possible to produce a unique general bio-

chem space [24] due to the “lack of invariance of the chemical space” [36]. The chemical space will depend on the particular molecular representation (*vide supra*) as well as the different ways to compute molecular similarity [35, 72]. What that being said, several approaches have been used to represent biochem, druglike, pharmacological and related chemical spaces on a case by case basis. Representative approaches with specific applications are discussed below.

### Druglike, Pharmacological and Other Focused Spaces

#### ChemGPS and ChemGPS-NP

The chemical global positioning system or ChemGPS is an approach developed to represent and navigate through *druglike* [73]; pharmacokinetically [74] and biologically relevant [39] chemical space. The approach is analogous to Mercator convention in geography. In ChemGPS or its variant ChemGPS-NP [39], the chemical space is constructed using a set of objects that are represented by chemical structures. The objects include a set of satellite structures and a set of core structures. The satellite structures are molecules with extreme values (e.g., at least one property value) that are located intentionally outside the chemical space establishing the limits of the chemical space under study. The notion of satellite structures can be related to the *abstract molecular basis vectors* proposed recently to represent chemical space [37]. The core structures maintain the balance of the model filling the core of the chemical space [73]. In addition to the objects a set of rules are also required. The set of rules are associated with the descriptors and are provided by the principal properties derived with PCA. To construct the *druglike chemical space* or *drug space* [73] the set of descriptors are associated with drug like characteristics including properties related to the size, lipophilicity, polarizability, charge, flexibility, rigidity, and hydrogen bond capacity [73]. Following this approach, a pharmacokinetically relevant space was constructed considering 18 VolSurf descriptors associated with ADME properties [74]. For ChemGPS-NP [39] a total of 35 descriptors including MW, TPSA, AlogP and several other counts of atoms and topology related properties were used. In each case, a chemical space was constructed for a training set: the ChemGPS data set for the drug and pharmacokinetically relevant spaces were composed of 423 compounds [73, 74] and the ChemGPS-NP was formed with 1,779 molecules [39]. The corresponding ChemGPS and ChemGPS-NP were used to estimate, *via* PCA score predictions, the properties of test sets. Thus, the ChemGPS corresponding to drug space was employed to classify 22,000 compounds with few outliers. The ChemGPS-NP was used to predict 619,382 compounds with no outliers. The ChemGPS approach is applied in *chemography*, defined as “the art of navigating the chemical space” [73, 75]. The ChemGPS and ChemGPS-NP are intended to be reference systems to compare different sets of molecules [73, 39].

#### Pharmacological Space

The pharmacological space of several different targets included in the MDL Drug Data Report (MDDR) has been represented as a branched tree obtained with a minimal spanning tree [47, 48] and as molecular property spaces [76]. Related drug-target networks including known FDA-approved drugs collected in the DrugBank database have

also been published recently [77]. Several other applications of target-based networks are reviewed in ref. [7].

#### Receptor-Relevant Subspace

Pearlman and Smith [49] developed the concept of receptor-relevant subspace in which the dimensions of the chemical space are defined such as the active compounds for that receptor are tightly clustered. Molecules were represented using BCUT descriptors and the dimensionality reduction was carried out with the algorithm *activity-seeded structure-based clustering* developed by the same authors [49]. As an example, 74 active acetylcholinesterase (AChE) inhibitors were visualized in the corresponding AChE-receptor-relevant subspace within the chemistry space of 70,000 MDDR compounds. Two virtual combinatorial libraries were also visualized in the same graph [49]. Representing the chemical space with metrics relevant to the AChE receptor produces that active AChE inhibitors are near neighbors of each other. Pearlman and Smith showed that they were unable to cluster active compounds in the chemical space if this is constructed using irrelevant metrics to the AChE receptor. As a consequence, active compounds could be missed in virtual screening based on a chemical space designed with irrelevant metrics.

#### Binding Site-Based Chemical Space

A visualization of the so-called binding site-based chemical space of structures from the PDB has been generated using PCA [78]. In this study, structures in complex with small molecules containing carboxylate groups (176 binding sites), 180 sulfonate moieties and 101 phosphonate groups were analyzed. Binding sites were represented using experimental and theoretical descriptors including frequency distribution of residue types, formal charge, hydrogen bonding, number of water molecules, hydrophobicity, flexibility and bulkiness. The binding site-based chemical space was used to analyze the bioisosteric relationships among the carboxylic, sulfonic and phosphonic groups [78].

#### Therapeutic and ADMET-Related Spaces

The comparison of compound collections targeted to specific therapeutic areas has been of interest in drug design projects. Examples of such databases reviewed ahead include GPCR ligands, molecules used in the treatment of cancer and compounds with anti-AIDS activity. Visualization of compound databases related to ADMET properties such as orally active, toxic and metabolites are also presented.

#### GPCR and Non-GPCR Ligands

The biochem space of a proprietary collection with more than 3,000 GPCR ligands, directed to 130 different receptors, was compared with a diverse set of the same size not targeted to GPCRs and obtained at random from a commercial database. Molecules were represented using a dictionary-based fingerprint and a topological pharmacophore point histogram descriptor [24]. The similarity of the molecules was measured with the inverse Tanimoto coefficient. The biochem space was visualized with PCA and SOMs. In this study both visualization techniques were compared in terms of the ability to distinguish between GPCR from non-GPCR ligands. SOMs showed a superior performance over PCA given the molecular representations and conditions of the study [24].

In a separate study, 11,800 compounds screened across four GPCR targets and one kinase target, were visualized by means of GTM and hierarchical GTM [20]. Molecules were represented with 11 whole-molecule physicochemical properties including molecular weight, molecular solubility, polar surface area and AlogP. Visualization using GTM and hierarchical GTM was compared with PCA and SOM. The GTM plots showed clearer clusters of compounds and were more informative than visualizations with PCA and SOM. This work applied to compounds screened against GPCR and kinase targets can be applied to other large data sets. The integration of the GTM approach with dynamic tools (Table 1) can further enhance the interpretation of the results [20].

### **Cancer Medicinal Chemistry Space**

Cancer medicinal chemistry space has been visualized using PCA [50] comparing the regions in the space covered by 12,714 compounds with known cancer activity, 36,683 known cancer inactives and 109,466 hit-like compounds obtained from the ZINC database. Compounds were represented with 48 two-dimensional molecular descriptors associated with the atomic nature, molecular size, polarity, lipophilicity and flexibility of the molecules. One of the major conclusions derived from the study is that anticancer compounds occupy a much wider area of the chemical space than the one covered by the hit-like molecules [50].

### **5-HT<sub>3</sub> Antagonists and HIV-1 Protease Inhibitors**

The chemical space of the 5-hydroxytryptamine 3 (5-HT<sub>3</sub>) receptor antagonists and HIV-1 protease inhibitors has been visualized using Sammon mapping, a non-linear mapping technique [79]. In the same study different structural representations were compared. The molecules were part of a virtual screening study involving 118,346 molecules from WOMBAT and 149,414 molecules from MDDR [79]. To construct the Sammon mapping, molecules were represented with Daylight fingerprints, topological autocorrelations descriptors and radial distribution function. Visualization of the chemical space was helpful for further interpreting the virtual screening results [79].

### **Estrogens, Monoamino Oxidase Inhibitors and Other Sets**

Xie *et al.* applied singular value decomposition as a low-dimensional projection method to represent the chemical space of four data sets with compounds in the range of 58 to 27,255 structures [51]. Datasets were formed by molecules with different types of biological activity, for example, estrogens, monoamino oxidase inhibitors and pesticides. The method was able to project similar structures close together in the two-dimensional maps [51].

### **NCI AIDS Antiviral Screen**

The data mining and visualization tool *ChemSpaceShuttle* [52] combines diverse algorithms for multi-dimension reduction including linear and non-linear projection techniques and SOMs [52]. The approach has been applied in compound selection and focused library design. As an example, the NCI AIDS collection was analyzed with *ChemSpaceShuttle* to visualize the separation between actives and inactives. The 29,184 molecules analyzed were represented using topological pharmacophore descriptors. A version of *ChemSpaceShuttle* is available on the Web (Table 1). SOMs have also been applied

to visualize the chemical space of approximately 42,000 compounds in the NCI AIDS antiviral screen. The SOM is available on the Web (Table 1).

### **Orally Active Compounds**

The chemical space of 9,114 orally active compounds from the MDDR-3D database has been visualized by means of tree-structured SOM (TS-SOM) which is a variation of a basic SOM [53]. TS-SOM includes a neighborhood function criterion and hierarchical clustering. Molecules were represented using VolSurf descriptors. Other techniques employed for visualization were Sammon mapping, U matrix and k-means clustering. In this study the four techniques provide similar results of clustering the orally active data set. Furthermore, the methods proved to effectively map a large number of compounds using physicochemical characteristics [53].

### **Toxic and Non-Toxic Molecules**

A visualization of the chemical space of toxic and non-toxic molecules was reported using SOMs [54]. In a series of SOMs, a total of 16,085 toxic compounds from the Registry of Toxic Effects of Chemical Substances were represented along with a set of 17,000 compounds obtained from the Investigational Drugs database. Four categories were represented by the toxic molecules namely mutagenic, tumorigenic, irritant and reproductive effective. Chemical structures were described in terms of proprietary molecular fingerprints and the similarity was measured with the Tanimoto coefficient. The SOMs were valuable for visually distinguishing between toxic and non-toxic compounds and were useful tools for further interpreting fragment-based statistical analysis and other classification studies performed in the same work [54].

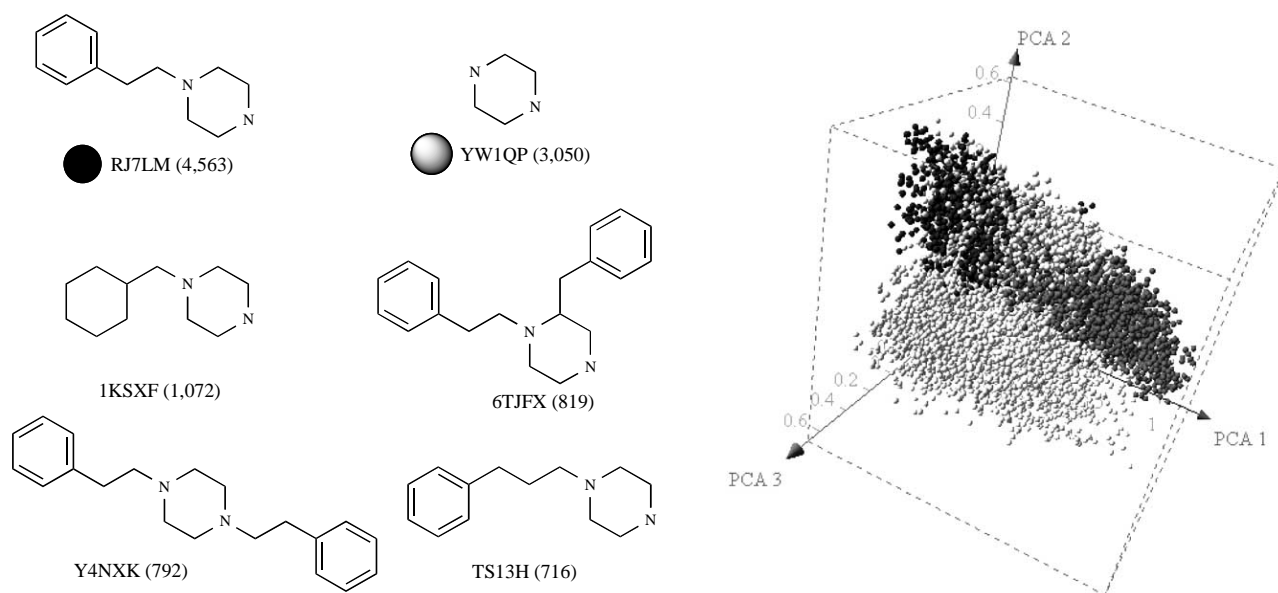
### **Metabolites and Non-Metabolites**

The chemical space of 1,811 metabolites and 4,598 non-metabolites has been visualized using SOMs [80]. Molecules were represented using radial distribution functions and global molecular descriptors (topological and physicochemical). It was observed that the global molecular descriptors were more accurate for distinguishing metabolites from non-metabolites [80].

## **STRUCTURAL ANALYSIS AND CHEMICAL SPACE**

In addition to the multi-dimensional data reduction techniques reviewed above, structural analysis has been useful for exploring the chemical space of compound collections. Substructural analysis in drug discovery has been reviewed by Villar *et al.* [81]. Applications of this technique includes a chemotype-based hierarchical classification of the NCI AIDS collection [82], structural classification of natural products [83], molecules assayed for pyruvate kinase activity and pesticides [84], and visualization of the molecules patented for activity on kinases [85].

Fig. (3) presents a recent substructural analysis of a combinatorial library, **II**, (Fig. 2) conducted by our group with the program Molecular Equivalence Indices (MEQI) [86]. The most representative cyclic systems, with a population of at least of 700 molecules, are displayed. In MEQI, cyclic systems (e.g., frameworks) are formed by removing the side chains (i.e., all vertices of degree one) from the initial



**Fig. (3).** Common cyclic systems (e.g., frameworks) of combinatorial library **II** (frequency indicated in parenthesis) and mapping of the two most populated frameworks into the chemical space obtained from shape-based 3D-similarity.

chemical graph [86]. Each cyclic system is assigned a code of five characters that uniquely identify that chemotype [87]. In Fig. (3) the number of compounds in the combinatorial library with the indicated cyclic system is shown in parenthesis. Interestingly, just six cyclic systems account for about one third of the entire library with 31,320 compounds (*vide supra*). The two most populated cyclic systems (i.e., chemotype identifiers RJ7LM and YW1QP) were further mapped into the chemical space of **II**, also depicted in Fig. (3). The chemical space was generated with the overlay of three-dimensional structures. To obtain this representation, 10 diverse molecules were selected with the MaxMin approach algorithm implemented in MOE [66] using MDL keys and the Tanimoto coefficient. A low energy conformation was selected for each of the 10 diverse molecules that was employed to perform a 3D-similarity analysis with the Rapid Overlay of Chemical Structures (ROCS) program [88]. The conformation library was obtained with OMEGA [89]. The comboscore similarities of each of the molecules in the library with each of the 10 diverse molecules formed a 10-dimensional matrix that was subject to PCA. The first three principal components represented in Fig. (3) explain 72.6% of the total variance. Noteworthy, molecules with the same cyclic system occupy a similar region in the chemical space. For example, molecules with the cyclic system RJ7LM all occupy a similar region in chemical space that is distinct from those molecules with the cyclic system YW1QP (Fig. 3). Similar conclusions can be obtained from mapping the chemical space of other cyclic systems in the library. Conclusions from the visual analysis suggest that molecular frameworks play an important role in the distribution of molecules in the chemical space and therefore can help to guide *expeditions through the chemical space*. Further investigation on the chemical space representation derived from shape-based 3D-similarity analysis, including the effect on the number and diversity of reference molecules, is being conducted in our group and will be reported in a separate work.

## CONCLUSION

Current definitions of chemical space and the general intuition of this concept usually refer to all possible molecules. Attempts to quantify the size of chemical space reveal that, similar to the cosmic universe, it could be infinite. In this context, a small number of molecules is currently known and it seems that an even smaller number of compounds have therapeutic interest. Among the plethora of computational techniques used to obtain information from large and/or multi-dimensional data like the ones generated in many current drug design projects, data visualization offers an intuitive approach for finding associations in complex data. Visualization of the chemical space has been useful for assessing the diversity of different data sets, exploring the relationships among collections and evaluating the potential for covering other regions in chemical space yet to be explored. Further applications include support in library selection and design, virtual screening and the development of algorithms to assess molecular representation and diversity. Visualization of the chemical space, however, is not a trivial task. The final representation will depend on the molecular description that typically leads to multi-dimensional spaces, and on the algorithm to reduce the dimensionality. The goal is to portray the original multi-dimensional data into two- or three-dimensions amenable to interpretation by the human eye. Investigation of novel ways to represent molecules and the development of algorithms, programs and approaches for visualization is the subject of intense research. To this extent, our group is currently developing approaches for representing chemical space using shape-based 3D-similarity as well as using MFS maps to study molecular diversity in combinatorial libraries.

## ACKNOWLEDGEMENTS

This work was supported by the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development. We are grateful to Dr. Mark Johnson, Pannanugget Consulting, LLC, for providing the pro-

gram MEQI and to OpenEye Scientific Software, Inc. for providing ROCS and OMEGA. Stimulating discussions with Dr. Gerald M. Maggiora are gratefully acknowledged as well as the assistance of Joseph Rios for enumerating libraries.

## ABBREVIATIONS

5-HT3	=	5-Hydroxytryptamine 3
AChE	=	Acetylcholinesterase
ADMET	=	Absorption, distribution, metabolism, excretion and toxicity
AIDS	=	Acquired immune deficiency syndrome
DOS	=	Diversity-oriented synthesis
FDA	=	U.S Food and Drug Administration
GPCR	=	G-protein-coupled receptor
GTM	=	Generative topographic mapping
LogP	=	Octanol/water partition coefficient
MDDR	=	MDL Drug Data Report
MEQI	=	Molecular equivalence index
MFS	=	Multi-fusion similarity
MOE	=	Molecular Operating Environment
MW	=	Molecular weight
NCI	=	National Cancer Institute
PCA	=	Principal component analysis
PDB	=	Protein Data Bank
ROCS	=	Rapid overlay of chemical structures
SOM	=	Self-organizing map
TPSA	=	Topological polar surface area
TS-SOM	=	Tree-structured self-organizing map
WOMBAT	=	World of molecular bioactivity

## REFERENCES

- Scior, T.; Bernard, P.; Medina-Franco, J.L.; Maggiora, G.M. *Mini-Rev. Med. Chem.*, **2007**, *7*, 851-860.
- Baker, M. *Nat. Rev. Drug Discov.*, **2006**, *5*, 707-708.
- Shelat, A.A.; Guy, R.K. *Curr. Opin. Chem. Biol.*, **2007**, *11*, 244-251.
- Balakin, K.V.; Savchuk, N.P. *Curr. Comput. Aided Drug Des.*, **2006**, *2*, 1-19.
- Bajorath, J. *Nat. Rev. Drug Discov.*, **2002**, *1*, 882-894.
- Root, D.E.; Kelley, B.P.; Stockwell, B.R. *Curr. Opin. Drug Discov. Dev.*, **2002**, *5*, 355-360.
- Ekins, S.; Mestres, J.; Testa, B. *Br. J. Pharmacol.*, **2007**, *152*, 9-20.
- Harper, G.; Pickett, S. D. *Drug Discov. Today*, **2006**, *11*, 694-699.
- Lipinski, C.; Hopkins, A. *Nature*, **2004**, *432*, 855-861.
- Dobson, C.M. Chemical space and biology. *Nature*, **2004**, *432*, 824-828.
- Nat. Rev. Drug Discov.*, **2004**, *3*, 375-375.
- In *Charting chemical space: finding new tools to explore biology* *Horizon Symposia* **2004** <http://www.nature.com/horizon/chemicalspace/index.html>.
- Geysen, H.M.; Schoenen, F.; Wagner, D.; Wagner, R. *Nat. Rev. Drug Discov.*, **2003**, *2*, 222-230.
- Bohacek, R.S.; McMartin, C.; Guida, W.C. *Med. Res. Rev.*, **1996**, *16*, 3-50.
- Ertl, P. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 374-380.
- Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.*, **2000**, *28*, 235-242.
- Ogata, K.; Isomura, T.; Yamashita, H.; Kubodera, H. *QSAR Comb. Sci.*, **2007**, *26*, 596-607.
- Fink, T.; Reymond, J.-L. *J. Chem. Inf. Model.*, **2007**, *47*, 342-353.
- Howe, T. J.; Mahieu, G.; Marichal, P.; Tabruyn, T.; Vugts, P. *Drug Discov. Today*, **2007**, *12*, 45-53.
- Maniyar, D. M.; Nabney, I. T.; Williams, B. S.; Sewing, A. *J. Chem. Inf. Model.*, **2006**, *46*, 1806-1818.
- Oellien, F.; Ihlenfeldt, W. D.; Gasteiger, J. *J. Chem. Inf. Model.*, **2005**, *45*, 1456-1467.
- Ahlberg, C. *Drug Discov. Today*, **1999**, *4*, 370-376.
- Shi, L. M.; Fan, Y.; Lee, J. K.; Waltham, M.; Andrews, D. T.; Scherf, U.; Paull, K. D.; Weinstein, J. N. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 367-379.
- vonKorff, M.; Hilpert, K. *J. Chem. Inf. Model.*, **2006**, *46*, 1580-1587.
- Schnur, D. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 36-45.
- Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. *Chem. Biol. Drug Des.*, **2007**, *70*, 393-412.
- Smellie, A.; Wilson, C. J.; Ng, S. C. *J. Chem. Inf. Model.*, **2006**, *46*, 201-207.
- Agrafiotis, D. K.; Shemanarev, M.; Connolly, P. J.; Farnum, M.; Lobanov, V. S. *J. Med. Chem.*, **2007**, *50*, 5926-5937.
- Amari, S.; Aizawa, M.; Zhang, J.; Fukuzawa, K.; Mochizuki, Y.; Iwasawa, Y.; Nakata, K.; Chuman, H.; Nakano, T. *J. Chem. Inf. Model.*, **2006**, *46*, 221-230.
- Stoner, C. L.; Gifford, E.; Stankovic, C.; Lepsey, C. S.; Brodfuehrer, J.; Prasad, J.; Surendran, N. *J. Pharm. Sci.*, **2004**, *93*, 1131-1141.
- Liu, T. Q.; Lin, Y. M.; Wen, X.; Jorissen, R. N.; Gilson, M. K. *Nucleic Acids Res.*, **2007**, *35*, D198-D201.
- LigPrep, version 2.1, Schrödinger, LLC, New York, NY, 2005.
- QikProp, version 3.0, Schrödinger, LLC, New York, NY, 2007.
- Spotfire, version 9.1.1, TIBCO Software, Inc., Somerville, MA. Available at: <http://www.spotfire.tibco.com>.
- Sheridan, R. P.; Kearsley, S. K. *Drug Discov. Today*, **2002**, *7*, 903-911.
- Maggiora, G. M. *J. Chem. Inf. Model.*, **2006**, *46*, 1535-1535.
- Raghavendra, A. S.; Maggiora, G. M. *J. Chem. Inf. Model.*, **2007**, *47*, 1328-1340.
- Jolliffe, I. T. *Principal Component Analysis* Second ed.; Springer: New York, **2002**.
- Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. *J. Nat. Prod.*, **2007**, *70*, 789-794.
- Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, **1997**.
- Terfloth, L.; Gasteiger, J. *Drug Discov. Today*, **2001**, *6*, 102-108.
- Manallack, D. T.; Livingstone, D. J. *Eur. J. Med. Chem.*, **1999**, *34*, 195-208.
- Yan, A. X. *Comb. Chem. High Throughput Screen.*, **2006**, *9*, 473-480.
- Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 218-227.
- Sadowski, J.; Wagener, M.; Gasteiger, J. *Angew. Chem., Int. Ed. in English*, **1996**, *34*, 2674-2677.
- Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Second ed. Weinheim, **1999**.
- Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K. *Nat. Biotechnol.*, **2007**, *25*, 197-206.
- Hert, J.; Keiser, M. J.; Irwin, J. J.; Oprea, T. I.; Shoichet, B. K. *J. Chem. Inf. Model.*, **2008**, *48*, 755-765.
- Pearlman, R. S.; Smith, K. M. *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 28-35.
- Lloyd, D. G.; Golfis, G.; Knox, A. J. S.; Fayne, D.; Meegan, M. J.; Oprea, T. I. *Drug Discov. Today*, **2006**, *11*, 149-159.
- Xie, D.; Tropsha, A.; Schlick, T. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 167-177.
- Givehchi, A.; Dietrich, A.; Wrede, P.; Schneider, G. *QSAR Comb. Sci.*, **2003**, *22*, 549-559.
- Matero, S.; Lahtela-Kakkonen, M.; Korhonen, O.; Ketolainen, J.; Lappalainen, R.; Poso, A. *Chemom. Intell. Lab. Syst.*, **2006**, *84*, 134-141.
- vonKorff, M.; Sander, T. *J. Chem. Inf. Model.*, **2006**, *46*, 536-544.
- Lee, M. L.; Schneider, G. *J. Comb. Chem.*, **2001**, *3*, 284-289.
- Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.*, **1997**, *23*, 3-25.
- Houghten, R. A.; Pinilla, C.; Giulianotti, M. A.; Appel, J. R.; Dooly, C. T.; Nefzi, A.; Ostresh, J. M.; Yu, Y. P.; Maggiora, G. M.;

- Medina-Franco, J. L.; Brunner, D.; Schneider, J. J. *Comb. Chem.*, **2008**, *10*, 3-19.
- [58] Fink, T.; Bruggesser, H.; Reymond, J.-L. *Angew. Chem., Int. Ed.*, **2005**, *44*, 1504-1508.
- [59] van Deursen, R.; Reymond, J.-L. *ChemMedChem*, **2007**, *2*, 636-640.
- [60] Haggarty, S. J.; Clemons, P. A.; Wong, J. C.; Schreiber, S. L. *Comb. Chem. High Throughput Screen.*, **2004**, *7*, 669-676.
- [61] Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. *Mol. Divers.*, **1996**, *2*, 64-74.
- [62] Martínez-Mayorga, K.; Medina-Franco, J. L.; Giulianotti, M. A.; Pinilla, C.; Dooley, C. T.; Appel, J. R.; Houghten, R. A. *Bioorg. Med. Chem.*, **2008**, *16*, 5932-5938.
- [63] Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. *Nucleic Acids Res.*, **2008**, *36*, D901-D906.
- [64] Nefzi, A.; Ostresh, J. M.; Yu, J.; Houghten, R. A. *J. Org. Chem.*, **2004**, *69*, 3603-3609.
- [65] MDL Information Systems Inc., San Leandro, CA. Available at: <http://www.mdli.com>.
- [66] Molecular Operating Environment [MOE] 2007, Chemical Computing Group Inc., Montreal, Quebec, Canada. Available at: <http://www.chemcomp.com> 2007.
- [67] Willett, P.; Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.*, **1998**, *38*, 983-996.
- [68] Fitzgerald, S. H.; Sabat, M.; Geysen, H. M. *J. Chem. Inf. Model.*, **2006**, *46*, 1588-1597.
- [69] Fitzgerald, S. H.; Sabat, M.; Geysen, H. M. *J. Comb. Chem.*, **2007**, *9*, 724-734.
- [70] Johnson, M. A.; Maggiora, G. M. *Concepts and Applications of Molecular Similarity*; Wiley: New York, **1990**.
- [71] Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. *J. Med. Chem.*, **2002**, *45*, 4350-4358.
- [72] Martin, Y. C. *J. Comb. Chem.*, **2001**, *3*, 231-250.
- [73] Oprea, T. I.; Gottfries, J. J. *Comb. Chem.*, **2001**, *3*, 157-166.
- [74] Oprea, T. I.; Zamora, I.; Ungell, A. L. *J. Comb. Chem.*, **2002**, *4*, 258-266.
- [75] Oprea, T. I. *Curr. Opin. Chem. Biol.*, **2002**, *6*, 384-389.
- [76] Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. *Nat. Biotechnol.*, **2006**, *24*, 805-815.
- [77] Yildirim, M. A.; Goh, K. I.; Cusick, M. E.; Barabasi, A. L.; Vidal, M. *Nat. Biotechnol.*, **2007**, *25*, 1119-1126.
- [78] Macchiarulo, A.; Pellicciari, R. *J. Mol. Graphics Model.*, **2007**, *26*, 728-739.
- [79] Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. J. *Comput. Aided Mol. Des.*, **2007**, *21*, 617-640.
- [80] Gupta, S.; Aires-De-Sousa, J. *Mol. Divers.*, **2007**, *11*, 23-36.
- [81] Villar, H. O.; Hansen, M. R.; Kho, R. *Curr. Comput. Aided Drug Des.*, **2007**, *3*, 59-67.
- [82] Medina-Franco, J. L.; Petit, J.; Maggiora, G. M. *Chem. Biol. Drug Des.*, **2006**, *67*, 395-408.
- [83] Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. *Proc. Natl. Acad. Sci. USA*, **2005**, *102*, 17272-17277.
- [84] Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. *J. Chem. Inf. Model.*, **2007**, *47*, 47-58.
- [85] Southall, N. T.; Ajay J. *Med. Chem.*, **2006**, *49*, 2103-2109.
- [86] Xu, Y. J.; Johnson, M. *J. Chem. Inf. Comput. Sci.*, **2002**, *42*, 912-926.
- [87] Xu, Y.; Johnson, M. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 181-185.
- [88] ROCS, version 2.3.1, OpenEye Scientific Software Inc., Santa Fe, NM. Available at: <http://www.eyesopen.com>.
- [89] OMEGA, version 2.2.1, OpenEye Scientific Software Inc., Santa Fe, NM. Available at: <http://www.eyesopen.com>.