

# Chemical space networks: a powerful new paradigm for the description of chemical space

Gerald M. Maggiora · Jürgen Bajorath

Received: 14 April 2014 / Accepted: 4 June 2014 / Published online: 13 June 2014  
© Springer International Publishing Switzerland 2014

**Abstract** The concept of chemical space is playing an increasingly important role in many areas of chemical research, especially medicinal chemistry and chemical biology. It is generally conceived as consisting of numerous compound clusters of varying sizes scattered throughout the space in much the same way as galaxies of stars inhabit our universe. A number of issues associated with this coordinate-based representation are discussed. Not the least of which is the continuous nature of the space, a feature not entirely compatible with the inherently discrete nature of chemical space. Cell-based representations, which are derived from coordinate-based spaces, have also been developed that facilitate a number of chemical informatic activities (e.g., diverse subset selection, filling ‘diversity voids’, and comparing compound collections). These representations generally suffer the ‘curse of dimensionality’. In this work, networks are proposed as an attractive paradigm for representing chemical space since they circumvent many of the issues associated with coordinate- and cell-based representations, including the curse of dimensionality. In addition,

their relational structure is entirely compatible with the intrinsic nature of chemical space. A description of the features of these chemical space networks is presented that emphasizes their statistical characteristics and indicates how they are related to various types of network topologies that exhibit random, scale-free, and/or ‘small world’ properties.

**Keywords** Chemical space · Molecular representations · Descriptor vectors · Cell-based methods · Molecular networks · Chemical space networks

## Introduction

Although the notion of chemical space is a relatively new one, it has provided an excellent framework for systematically characterizing many aspects of chemical behavior, especially in the areas of chemical informatics, medicinal chemistry, and drug discovery. And it is beginning to creep into other areas of chemistry such as materials research [1]. This raises the question of whether current methods for representing chemical spaces are optimal in meeting evolving research needs in the chemical sciences. Here it is argued that a new representation based on the network paradigm can augment current procedures and provide added benefits. In other words, ‘molecular networks’ are thought to provide an alternative way to represent and navigate chemical space that complements and further extends current views of chemical space. To set the stage for the following discussion, it should be noted that chemical space exploration is, first and foremost, a computational task and is often motivated by the need to better understand structure–property relationships of small organic compounds. As a consequence, alternative representations of chemical space have been developed by

---

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

G. M. Maggiora  
Translational Genomics Research Institute, 445 North Fifth  
Street, Phoenix, AZ 85004, USA

J. Bajorath (✉)  
Department of Life Science Informatics, B-IT, LIMES Program  
Unit Chemical Biology and Medicinal Chemistry, Rheinische  
Freidrich-Wilhelms-Universität, Dahlmannstrasse 2,  
53113 Bonn, Germany  
e-mail: bajorath@bit.uni-bonn.de

considering the conditions and computational requirements for space navigation.

### Representations of chemical space

Typically, chemical space is envisioned as a set of ‘molecular points’ distributed within a multi-dimensional space much as the stars and planets are dispersed throughout our universe [2]. Although this coordinate-based model provides an appealing and computationally accessible representation of chemical space, it suffers from a number of defects (not necessarily in order of importance):

1. The space of points associated with coordinate-based representations is a priori continuous, but chemical space is inherently discrete since the potential number of molecules, though enormous, is nonetheless finite.
2. Chemical spaces are *not invariant* to the representations used to encode molecular information; different representations lead to different spaces and chemical neighborhood relationships may or may not be maintained among these representations.
3. Computed chemical spaces are generally of very high dimension and hence suffer from the ‘curse of dimensionality’ [3]. Such spaces are also subject to idiosyncratic behaviors that can be difficult to comprehend [4]. Hence, some form of dimensionality reduction is generally required for their effective use and interpretation.
4. The continuous components of vector representations generally have different units that need to be scaled to ensure they are compatible.
5. In many cases, the components of representation vectors, such as molecular fingerprints, have binary or categorical values that are incompatible with continuous chemical spaces. Hence, these vectors must often be converted to coordinate systems that are generally of much lower dimension resulting in a loss of information that may or may not be critical to further analyses.

Alternatively, a cell-based model of chemical space [5] can be obtained by superimposing a set of non-intersecting hypercubic cells that cover the associated coordinate-based chemical space. This creates a coarse-grained model since more than one molecule can inhabit a given cell—in set theoretic terms, it induces a partition of chemical space where each cell is an *equivalence class*. The location of cell boundaries may cause problems for nearby molecules since small changes in position proximal to cell boundaries can lead to significant changes in cell occupancies. Moreover, as the cellular location of a molecule depends on its coordinates, the inherently continuous nature of coordinate-based representations can

also influence the distribution of compounds data sets over discrete partitionings. Lastly, because cell-based spaces must typically be orthogonal and of low dimension to ensure meaningful analysis of compound distributions, dimensionality reduction is generally required, and the ensuing loss of information may also influence the final results. Although information is lost, the cell-based representation nevertheless facilitates carrying out many chemical informatic procedures such as comparing compound collections, selecting diverse subsets of compounds, and compound acquisition that can be more difficult to carry out solely within a coordinate-based framework.

Network-based representations provide a third alternative. Although the application of networks has seen explosive growth in numerous areas that include the World Wide Web, bioinformatics, social networks, citation networks, and airline networks, to name a few [6, 7], relatively little has been done in the chemistry field. The situation is beginning to change, however, and a number of network-related papers have been published at the interface of chemistry and closely related fields of biology, pharmacology, and medicine. The networks presented in these works generally involved relationships among macromolecules (e.g. protein–protein interaction networks) or ligands/drugs and their macromolecular targets (e.g. drug–target and target–target networks) [8–11]. None have focused exclusively on relationships of small molecules to one another.

This issue is, however, particularly germane to the analysis of ‘biologically relevant chemical space’, i.e., small regions of chemical space that contain compounds, including drugs, with specific biological activities, because networks are much easier to annotate with various molecular properties, including biological activity, than are other chemical space representations. For example, in a cell-based model of chemical space (vide supra), compound distributions and associated properties can only be viewed if the model is reduced to three dimensions. This can result in a substantial loss of information and hence often precludes quantitative assessment of relevant compound and property distributions.

### Characteristics of chemical space networks

Although chemical reaction networks have been studied for many years, only five papers published to date have applied the network paradigm to chemical spaces [12–16]. Thus, the field of chemical space networks (CSNs) is wide open to new and exciting possibilities, which begs the question as to why networks should be used to represent chemical spaces.

In addition to the annotation issue raised above, there are at least three formal reasons why CSNs provide a desirable representation of chemical space:

1. Most importantly, networks afford a ‘natural representation’ of chemical spaces since they capture the discrete structure of these spaces and the similarity, or distance, relationships between pairs of molecules residing within them—there is no need for the construction of a coordinate system or for any form of dimensionality reduction. Hence, they do not suffer the ‘curse of dimensionality’ as is the case with coordinate- and cell-based representations [3]. Interestingly, many networks appear to exhibit fractal dimensionality [17].
2. CSNs provide an appropriate conceptual framework for statistically analyzing many aspects of chemical spaces.
3. Given numerous applications of very large networks in other fields, efficient algorithms exist for analyzing many types of network features and for quantifying network characteristics [18]. In this regard, the field of social science has contributed significantly to their development and interpretation [7, 19].

However, it is important to recognize that issues associated with the non-invariance of chemical spaces to changes in molecular representation remain. In other words, changes associated with similarity relationships that are a consequence of utilizing alternative molecular representations and/or similarity functions or coefficients will also be reflected in a CSN. If this is not the case, and the network is insensitive to molecular details, it is unsuitable for chemical space representation.

Networks are large mathematical graphs and are made up of a set of vertices, also called nodes, and a set of edges connecting them.<sup>1</sup> The vertices of a CSN correspond to individual molecules. An edge is drawn between a pair of vertices if the similarity value or ‘distance’ of the corresponding molecular pair satisfies some threshold criterion. Molecular identifiers can label vertices and the edges connecting them can be labeled with various types of information such as similarity or distance values. The unlabeled graphs are called *threshold graphs*. For example, if the threshold similarity value chosen for a given molecular fingerprint representation is, say 0.85, edges will only be drawn between vertices associated with molecule pairs that have similarity values greater than or equal to 0.85.<sup>2</sup> Although unlabeled edges are simpler to handle than labeled

ones, information is lost, and thus, in some cases it is desirable to retain edge labeling at the expense of increased algorithmic complexity.<sup>3</sup>

Since similarity values and distances generally are symmetric, the corresponding edges are undirected, giving rise to undirected networks. In some cases, however, similarity may be treated asymmetrically, resulting in directed networks. Other types of directed networks can describe addition features of chemical spaces. For example, Guha and Van Drie [20] developed SALI, a structure–activity landscape index that can be used to identify putative cliffs in activity landscapes,<sup>4</sup> with larger SALI values generally being associated with steeper cliffs. They constructed directed graphs linking molecule pairs whose SALI values exceeded a given threshold and whose edges pointed towards the most active molecule in a cliff pair. Such graphs generally are made up of multiple relatively small subnetworks disconnected from one another. Krein and Sukumar [13] have also explored directed activity-cliff networks. However, it is important to point out that activity cliff networks can also be represented by undirected networks [16]. Figure 1a shows a section of such a global activity cliff network generated for currently available bioactive compounds that follows a different design idea. Here, nodes also represent compounds but edges indicate activity cliffs, i.e., they account not only for pairwise similarity but also for a potency difference large enough to qualify for cliff formation.

The pattern of network connectivity is called its *network topology*, and is designated by the symbol  $T(S_T)$  since it is directly related to the chosen similarity threshold value,  $S_T$ . A different network topology will most likely result if  $S_T$  is altered. Figure 1b illustrates different topologies of activity clusters within the global network. Thus, topology is an important parameter for characterizing networks both globally and locally. Figure 1c shows recurrent basic topologies identified in the activity cliff network and extensions of these topologies. In addition, in Fig. 1d, e, examples of activity cliff clusters with defined topology are shown including the compounds forming coordinated activity cliffs. The underlying structure–activity relationships would have been difficult to detect without the aid of the network representation. In CSNs, such local network features describe chemical neighborhoods, which are the primary focal point of structure–property analysis. The degree of connectivity of a network tends to increase as its similarity threshold is decreased until it ultimately

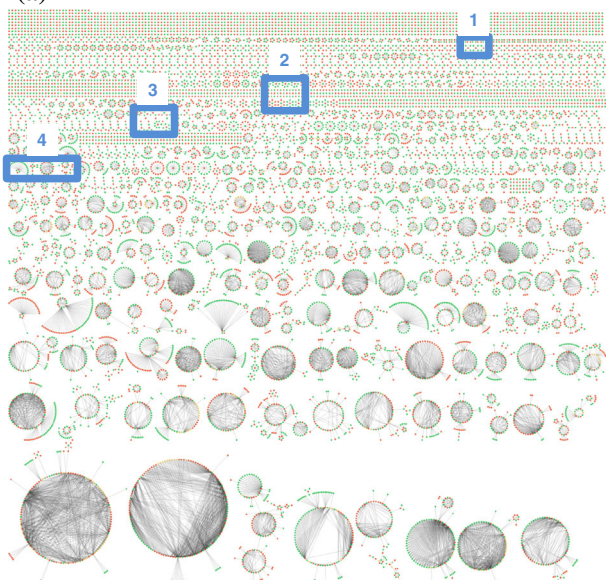
<sup>1</sup> When all vertices are connected to each other the network is called complete. Complete undirected networks with  $n$  vertices have  $n(n-1)/2$  edges.

<sup>2</sup> Similarity values that are strictly greater than a given threshold are called *proper* threshold values.


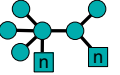

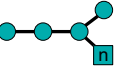




<sup>3</sup> Note that in rare cases the similarity between two dissimilar molecules may, nonetheless, be unity. This is due to limitations of the molecular representation that do not properly account for all the features that distinguish the two molecules from one another.

<sup>4</sup>  $SALI(i,j) = |\Delta Act(i,j)|/[1 - Sim(i,j)]$ , where  $\Delta Act(i,j)$  is the difference in the activities of a given compound pair  $(i,j)$ , and  $Sim(i,j)$  is their corresponding the similarity value.

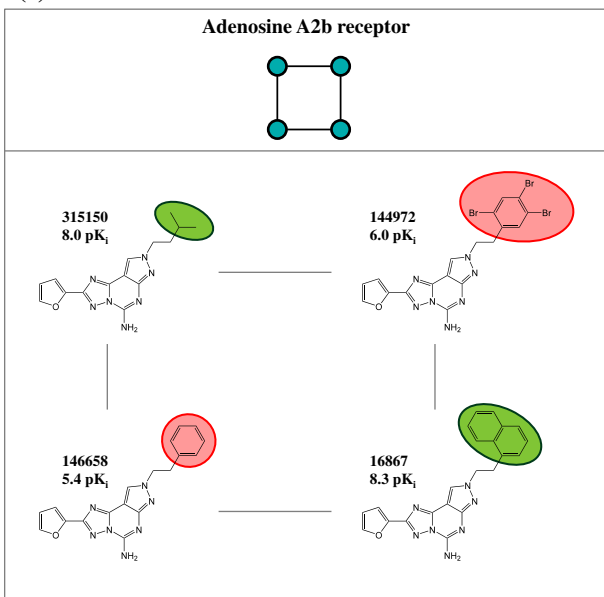
(a)



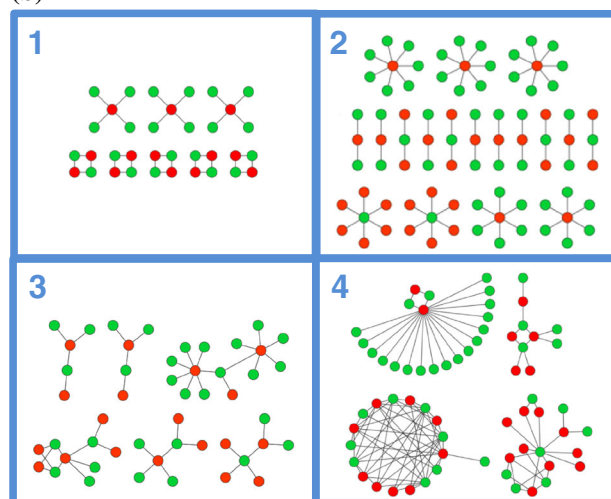
(c)

Main topology	Extensions of maintopology		
Star 	Twinstar 		
Chain 	Modified chain 		
Rectangle 	Modified rectangle 	Nested rectangle 	Fused rectangle 

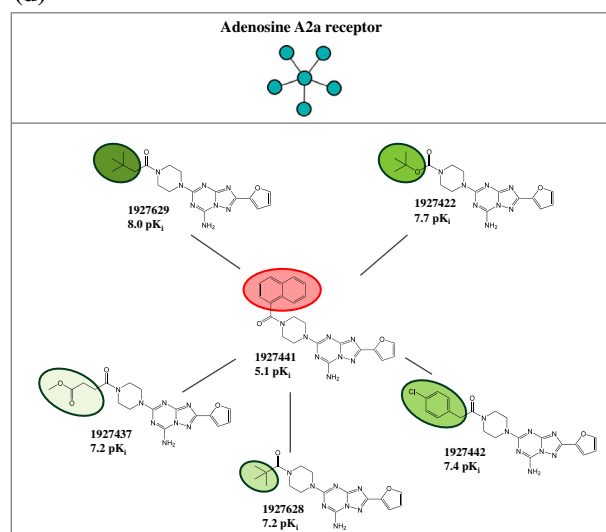
(e)



(b)



(d)





**Fig. 1** Activity cliff network. **a** A global activity cliff network derived from currently available active compounds is shown [16]. Vertices represent compounds and edges activity cliffs. Vertices of highly and weakly potent cliff partners are colored green and red, respectively. Small sections of the network containing exemplary activity cliff clusters are marked numbered. **b** These sections are enlarged and highlight activity cliff clusters of different size and topology. **c** The three most frequently observed activity cliff cluster topologies are schematically illustrated on the left and extensions of recurrent topologies are shown on the right. The three main topologies are termed *star*, *chain*, and *rectangle*, respectively. Squared nodes represent variable node numbers ( $n$ ). **d, e** Exemplary activity cliff clusters with defined topologies are shown and the compounds forming these clusters. Structural modifications of compounds are highlighted and colored according to compound activity (red low activity; green high activity). **d** A cluster with star topology formed by adenosine A2a receptor ligands and **e** a cluster with rectangle topology formed by adenosine A2b receptor ligands. The figure was adapted with permission from [16]. Copyright (2014) American Chemical Society

approaches zero, at which time every vertex becomes connected to every other vertex—such networks are called complete. Sampling networks at decreasing threshold values gives rise to an ensemble of threshold networks, each network being associated with a specific similarity (or distance) threshold value. Networks with lower similarity thresholds (higher distance thresholds) tend to be more connected, while the subnetworks corresponding to higher similarity (lower distance) threshold values are nested entirely within them, a situation reminiscent of that obtained in single-linkage agglomerative hierarchical clustering [21]. By contrast, the topology of edge-labeled networks, designated by the symbol  $T(S)$ , corresponds to a complete network where all of the vertices are connected since  $S > 0$  for all labeled edges. In some rare cases, highly dissimilar molecule pairs, which correspond to very weak links,  $S \approx 0$ , might be found.

Since chemical space is large so are the networks that represent them. Because of this a global graphical depiction cannot portray their detailed structure in any meaningful way. However, subsections of large networks that focus on selected chemical neighborhoods can often be displayed and graphically analyzed (cf. Figure 1). Matrix representations, which tend to be quite sparse, also do not provide a feasible alternative for very large networks.

Network features are generally analyzed statistically [18, 22]. The most important features are *vertex degree*, *degree correlations*, *local and global clustering coefficients*, *geodesic graph distances*, also called *shortest paths*, and *neighborhoods*. Since the topology of unlabeled CSNs changes in response to alterations in threshold value (vide supra), the statistics associated with different network features will likely change as well, and hence will functions of  $S_T$ .

Vertex degree, the most common feature, is simply the count of edges associated with a given vertex.<sup>5</sup> Degree

correlations, also called *assortativity*, represent correlations between the vertex degrees of adjacent (i.e. directly connected) vertices. Local and global clustering coefficients, as the name implies provide a measure of the degree that linked vertices tend to cluster together, and the geodesic or graph distances, represent the minimum number of edges connecting two vertices.<sup>6</sup> Unlike the situation in continuous spaces, network-based representations of *neighborhoods*, since they are discrete, are easier to understand and implement. For example, sophisticated limiting processes are not needed to describe the topology of network neighborhoods. The frequency or probability distributions associated with all of these network features afford a means for characterizing, albeit incompletely, network topologies and, hence, provide information on the characteristics of the associated chemical space.

An review by Albert and Barabási [23] provides an excellent overview of networks and describes a wide variety of their statistical and topological features (vide infra).

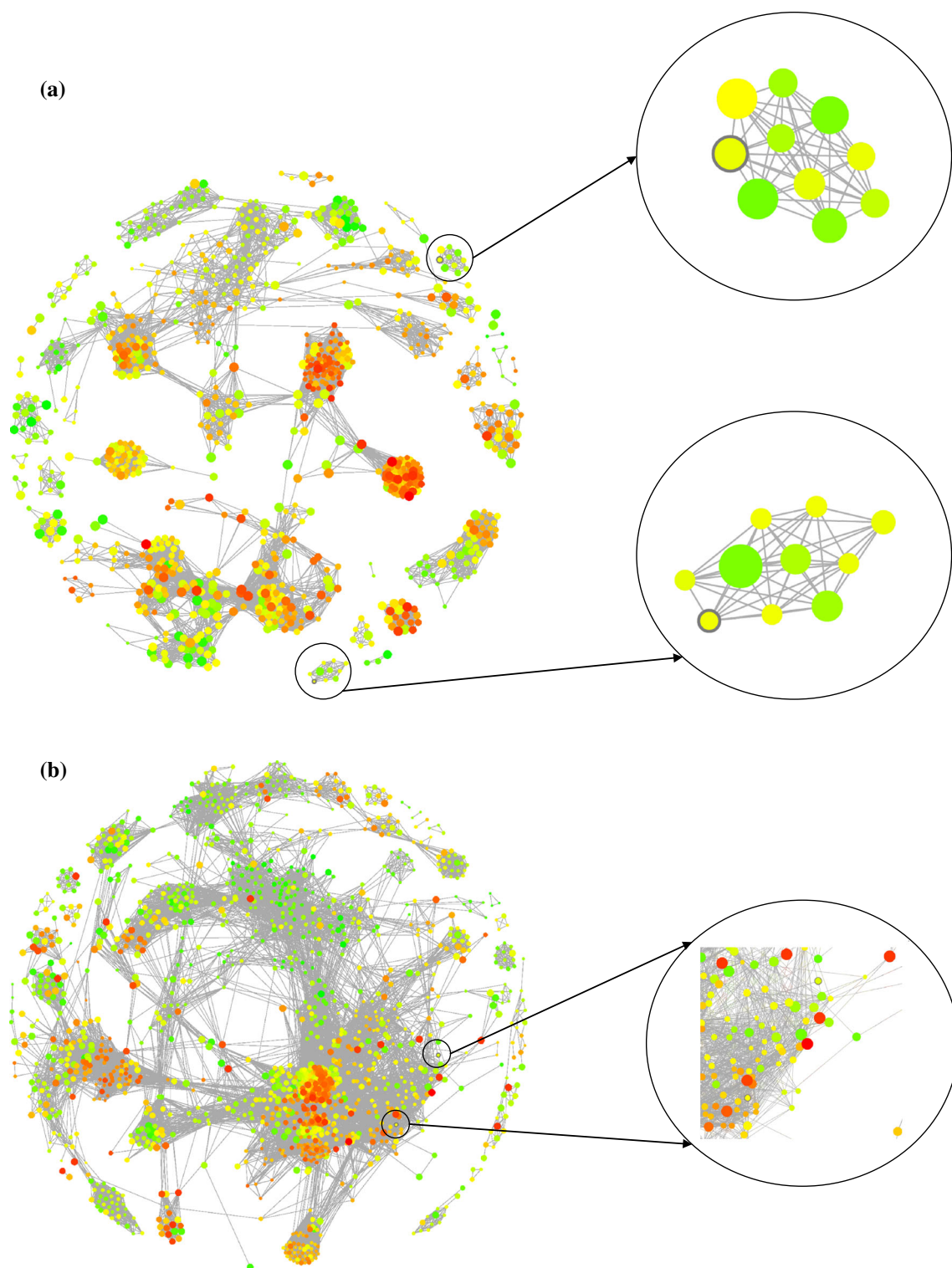
## Topologies of Chemical Space Networks

In networks with random connections between vertices, vertex degrees follow a Poisson distribution, which has an exponential tail [18, 22]. In such cases, vertex degrees do not deviate significantly from their average value. Other types of real world networks such as the World Wide Web, social networks, protein–protein interaction networks, citation networks, and airline networks [6, 7] tend, at least asymptotically, to follow *power law distributions* [18, 22]. Unlike Poisson distributions, power law distributions do not tail off as fast for large values of the vertex degree. Hence, highly connected vertices, generally called ‘hubs’, tend to be much more common in such networks. Networks that exhibit power law behavior are called *scale-free* [24].

Another interesting feature of many networks is their ‘small world’ character [25], which arises when the expected value of the geodesic or graph distance (i.e. shortest path) between a randomly selected pair of vertices in a network scales as  $\log(n)$ , with  $n$  being the number of vertices. The expression ‘small world behavior’, first given by Watts and Strogatz [26], was based on studies of social networks and derives from two of their features: *Homophily*, the desire to connect with others similar to ourselves, and *weak ties* that directly connect us with acquaintances that otherwise would be quite distant.

<sup>5</sup> In the case of directed graphs, the degree of incoming connections (‘in-degree’) is considered separately for that of the outgoing connections (‘out-degree’).

<sup>6</sup> Hence, shortest paths are integer valued. Note that it is possible that more than one shortest path exists between a given pair of vertices.



**Fig. 2** Activity landscapes. Shown are similarity-based compound network representations (so-called network-like similarity graphs [14]) for two compound activity classes: **a** 5-lipoxygenase inhibitors and **b** dopamine D3 receptor antagonists. Vertices are *color-coded* according to compound potency using a continuous *color* spectrum from *red* (high potency) to *yellow* and *green* (low potency) and *scaled*

in size according to their contribution to local SAR discontinuity [14, 15]. *Edges* indicate pairwise fingerprint similarity relationships. Vertices representing newly identified hits from ligand-based virtual screening are marked in the networks and regions containing these hits are encircled and enlarged. The figure was adapted with permission from [15]. Copyright (2011) American Chemical Society

Interestingly, power law and small world behaviors have been observed in several of the studies mentioned earlier [12, 13, 16]. Studies in the Baldi laboratory [27] also observed these chemical space behaviors, but their work focused primarily on molecular properties, structural features, and clustering, not on the relationship of individual compounds to one another.

Figure 2 shows similarity-based compound networks for activity classes into which ligand-based virtual screening hits are mapped [14]. Although not explicitly noted, as the networks were generated for SAR exploration, it provides a clear example of small world behavior with respect to the activities of individual targets. It also provides an exemplary depiction of what can be called *target promiscuity*, *aka scaffold hopping* [28] or *similarity cliffs* [29], where a single target exhibits activity for more than one class of compounds.

The form of the linked clusters of subnetworks depicted in Fig. 2 also suggests an interesting possible application for reduced graphs that heretofore have been applied primarily to individual molecules [30]. In reduced graphs, characteristic groups of connected atoms (e.g., functional groups) are condensed into single vertices significantly simplifying chemical graphs. In the present case, *reduced* CSNs could be constructed by condensing clusters of molecules into single vertices, thus affording a simpler depiction of the global inter-relationships among clusters of molecules.

## Conclusions

The question now arises as to whether the network paradigm can be applied in a way that facilitates many of the activities typically carried out in chemical space such as comparing compound collections, ligand-based virtual screening, compound acquisition, diversity assessment, and subset selection. While these issues have not been dealt with to date and, hence, are subjects for future research, the success of the network paradigm in elucidating many aspects of large networks such as is associated with the World Wide Web [31] suggests that a network-based representation of chemical spaces may provide an effective structure for their exploration, characterization, and exploitation.

An interesting development in this regard is the emergence of *graph databases* [32], which differ significantly from the typical relational databases in use today. Graph databases deliver powerful computational capabilities for processing large complex networks of highly linked information and, hence, should afford suitable means for treating CSNs as well.

As compound collections grow larger and become more diverse and complex in their biological annotations, the

need is clearly manifested for new methods that can effectively utilize the vast range of associated information. CSNs may provide another means that can help in accomplishing this goal. In any case, it seems reasonable to at least explore their capabilities to determine if, in fact, further development is warranted.

**Acknowledgments** The authors wish to thank Dr. Vijay Gokhale for reading the manuscript and for his helpful comments and Dr. Dagmar Stumpfe for the design of exemplary network representations and review of the manuscript.

## References

1. Workshop on Navigating chemical compound space for materials and bio design, held at the Institute for Pure and Applied Mathematics, University of California, Los Angeles, CA, March 14–June 17, 2011. <https://www.ipam.ucla.edu/programs/ccs2011/>. Accessed 3 April 2014
2. Dobson CM (2004) Chemical space and biology. *Nature* 432:824–828
3. Bellman RE (1961) Adaptive control processes. Princeton University Press, Princeton
4. Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley Publishing Company, Reading
5. Pearlman R, Smith K (2002) Novel software tools for chemical diversity. *3D QSAR Drug Design* 2:339–353
6. Barabási A-L (2003) Linked—how everything is connected to everything else and what it means for business, science, and everyday life. PLUME, Penguin Books, New York
7. Watts DJ (2003) Six degrees—the science of a connected age. W.W. Norton & Company, New York
8. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. *Nat Biotechnol* 24:805–815
9. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* 4:682–690
10. Keiser MJ, Roth BL, Armruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 25:197–206
11. Yildirim MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug-target network. *Nat Biotechnol* 25:1119–1126
12. Tanaka N, Ohno K, Niimi T, Moritomo A, Mori K, Orita M (2009) Small-world phenomena in chemical library networks: application to fragment-based drug discovery. *J Chem Inf Model* 49:2677–2686
13. Krein MP, Sukumar N (2011) Exploration of the topology of chemical spaces with network measures. *J Phys Chem A* 115:12905–12918
14. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J (2008) Structure-activity relationship anatomy by network-like similarity graphs and local structure-activity relationship indices. *J Med Chem* 51:6075–6084
15. Ripphausen P, Nisius B, Wawer M, Bajorath J (2011) Rationalizing the role of SAR tolerance for ligand-based virtual screening. *J Chem Inf Model* 51:837–842
16. Stumpfe D, Dimova D, Bajorath J (2014) Composition and topology of activity cliff clusters formed by bioactive compounds. *J Chem Inf Model* 54:451–461
17. Cohen R, Havlin S (2009) Scaling properties of complex networks and spanning trees. In: Bollobás B, Kozma R, Miklós (eds) Handbook of large-scale random networks. Springer, New York, pp 143–169

18. Newman MEJ (2010) *Networks—an introduction*. Oxford University Press Inc., New York
19. Wasserman S, Faust K (1994) *Social network analysis—methods and applications*. Cambridge University Press, Cambridge
20. Guha R, Van Drie JH (2008) Structure-activity landscape index: identifying and quantifying activity cliffs. *J Chem Inf Model* 48:646–658
21. Jain AK, Dubes RC (1988) *Algorithms for clustering data*. Prentice-Hall, Engelwood Cliffs
22. Kolaczyk ED (2009) *Statistical analysis of network data—methods and models*. Springer, New York
23. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
24. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
25. Watts DJ (1999) *Small worlds—the dynamics of networks between order and randomness*. Princeton University Press, Princeton
26. Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small world’ networks. *Nature* 393:440–442
27. Benz RW, Swamidass SJ, Baldi P (2008) Discovery of power-laws in chemical space. *J Chem Inf Model* 48:1138–1151
28. Schneider G, Neidhart W, Giller T, Schmid G (1999) ‘Scaffold hopping’ by topological pharmacophore search: a contribution to virtual screening. *Angew Chem Int Ed* 38:2894–2896
29. Iyer P, Stumpfe D, Vogt M, Bajorath J, Maggiora GM (2013) Activity landscapes, information theory, and structure-activity relationships. *Mol Inf* 32:421–430
30. Birchall K, Gillet VJ (2011) Reduced graphs and their applications in chemoinformatics, chapter 8. In: Bajorath J (ed) *Chemoinformatics and computational chemical biology*. Springer, New York, pp 197–212
31. Liu B (2011) *Web data mining: exploring hyperlinks, contents, and usage data*. Springer, Heidelberg
32. Robinson I, Webber J, Elfreem E (2013) *Graph databases*. O’Reilly Media Inc., Sebastopol